METHODS FOR THE DETERMINATION OF PROTEIN THREE-DIMENSIONAL STRUCTURE EMPLOYING HYDROGEN EXCHANGE ANALYSIS TO REFINE COMPUTATIONAL STRUCTURE PREDICTION

FIELD OF THE INVENTION

[0001] The present invention relates to methods for determining polypeptide and protein three-dimensional structures. In a particular aspect, the invention relates to methods for three-dimensional structure determination that employ hydrogen exchange analysis to refine, constrain and improve computational protein structure predictive methods.

BACKGROUND OF THE INVENTION

[0002] Considerable experimental work and time are required to precisely characterize the structure of a polypeptide of interest. In general, the techniques that are the easiest to use and which give the quickest answers, result in an inexact and only approximate idea of the nature of the critical structural features. Techniques in this category include the study of proteolytically generated fragments of the protein which retain binding function; recombinant DNA techniques, in which proteins are constructed with altered amino acid sequence (for example, by site-directed mutagenesis); epitope scanning peptide studies (construction of a large number of small peptides representing subregions of the intact protein followed by study of the ability of the peptides to inhibit binding of the ligand to receptor); covalent crosslinking of the protein to its binding partner in the area of the binding site, followed by fragmentation of the protein and identification of cross-linked fragments; and affinity labeling of regions of the receptor which are located near the ligand binding site of the receptor, followed by characterization of such "nearest neighbor" peptides.

[0003] Other techniques that are capable of finely characterizing polypeptide three-dimensional structure are considerably more difficult in practice. The most definitive techniques for the characterization of polypeptide structure, and receptor binding sites in particular, have been NMR spectroscopy and X-ray crystallography. While these techniques can ideally provide a precise characterization of relevant structural features, they have major

limitations, including inordinate amounts of time required for study, inability to study large proteins, and, for X-ray analysis, the need for protein and/or protein-binding partner crystals.

Peptide Amide Hydrogen Exchange

[0004] For more than 40 years, peptide amide hydrogen exchange techniques have been employed to study the thermodynamics of protein conformational change and the mechanisms of protein folding (Englander, et al. Methods Enzymol. 232:26-42 1994, Bai, et al. Methods Enzymol. 259:344 1995). More recently, they have proven to be increasingly powerful methods by which protein dynamics, domain structure, regional stability and function can be studied (Englander, et al. Protein Science 6:1101-9 1997 , Engen, et al. Analytical Chemistry 73:256A-265A 2001). Peptide amide hydrogens are not permanently attached to a protein, but continuously and reversibly interchange with hydrogen present in water. The chemical mechanisms of the exchange reactions are understood, and several well-defined factors can profoundly alter exchange rates. (Englander, et al. Methods Enzymol. 232:26-42 1994, Englander, et al. Anal. Biochem. 147:234-244 1985, Englander, et al. Methods Enzymol. 26:406-413 1972, Englander, et al. Methods Enzymol. 49G:24-39 1978) One of these factors is the extent to which a particular exchangeable hydrogen is exposed (accessible) to water.

[0005] Fully solvated amides. Peptide amide hydrogens that are freely accessible to water exchange at their maximal possible rate, with an average half-life of exchange of approximately one second at 0 °C and pH 7.0. (Molday, et al. Biochemistry 11:150 1972, Bai, et al. Proteins: Structure, Function, and Genetics 17:74-86 1993). The precise rate of exchange of a particular fully-solvated amide can vary more than thirty-fold from this average rate, depending upon the identity of the two amino acids flanking the amide bond. Exact exchange rates expected for fully solvent-exposed amide hydrogens can be reliably calculated from knowledge of the temperature, pH and primary amino acid sequence involved. (Molday, et al. Biochemistry 11:150 1972, Bai, et al. Proteins: Structure, Function, and Genetics 17:74-86 1993). When all amides in a substantial stretch of primary sequence exchange at the maximal rate, it usually indicates that the sequence is unstructured. High-throughput hydrogen exchange techniques can be used to quickly identify and localize such rapidly exchanging unstructured stretches of sequence within otherwise well-structured

proteins, and have demonstrated that truncated protein constructs depleted of such unstructured regions exhibit superior crystallization properties (see, for example, Examples herein). The same method also rapidly identifies and localizes the structured, but fully solvated, surface amides in proteins, and that this information can be used to dramatically refine structural predictions in a high throughput manner.

[0006] Partially solvated amides. In a structured protein, most peptide amide hydrogens exchange slower (up to 10^9-fold slower) than the maximal, fully solvated exchange rate, as they are not efficiently exposed to solvent water most of the time. Protein structure is not static, but is best considered as an ensemble of transiently unfolded states-the native state ensemble. Amide hydrogen exchange occurs only when a particular transient unfolding event fully exposes an amide to solvent. The ratio of exchange rates for a particular amide hydrogen, in the folded *vs.* random coil states is referred to as the exchange protection factor, and directly reflects the free energy change in the atomic environment of that particular hydrogen between unstructured and structured states of the protein. In this sense, amide hydrogens can be treated as atomic-scale sensors of highly localized free energy change throughout a protein and the magnitude of free energy change reported from each of a protein's amides in a folded *vs.* unfolded state is precisely equal to -RT ln (protection factor) (Bai, et al. Methods Enzymol. 259:344 1995). In effect, each peptide amide's exchange rate in a folded protein (when measured) directly and precisely reports the protein's structure and thermodynamic stability at the individual amino acid scale (Englander, et al. Methods Enzymol. 232:26-42 1994, Bai, et al. Methods Enzymol. 259:344 1995). Perhaps the single most important element of the instant invention's approach to structure determination is that this aggregate exchange rate data for a protein is treated as a "fingerprint" that is uniquely linked with its structure.

High Resolution, High Throughput Peptide Amide Hydrogen/ Deuterium Exchange-Mass Spectrometry (DXMS)

[0007] Deuterium exchange methodologies coupled with Liquid Chromatography Mass Spectrometry (LCMS), developed over the past 10 years, presently provide the most effective approach to study proteins larger than 30 kDa in size (Engen, et al. Analytical Chemistry 73:256A-265A 2001). Proteolytic and/or collision-induced dissociation (CID) fragmentation methods allow exchange behavior to be mapped to subregions of the protein(Engen, et al. Analytical Chemistry 73:256A-265A 2001, Hoofnagle, et al. Proceedings, National Academy of Sciences 98:956-961 2001, Resing, et al. J. Am Soc Mass Spectrom 10:685-702 1999, Mandell, et al. Anal. Chem. 70:39487-3995 1998, Mandell, et al. Proc Natl Acad Sci U S A 95:14705-10. 1998, Mandell, et al. J. Mol. Biol. 306:575-589 2001, Kim, et al. J Am Chem Soc 123:9860-6. 2001, Kim, et al. Biochemistry 40:14413-21. 2001, Zhang, et al. Protein Sci 10:2336-45. 2001, Kim, et al. Protein Sci 11:1320-9. 2002, Peterson, et al. Biochem J 362:173-81. 2002, Yan, et al. Protein Sci 11:2113-24. 2002). Building upon the pioneering work of Walter Englander, and David Smith (Englander, et al. Protein Science 6:1101-9 1997, Engen, et al. Analytical Chemistry 73:256A-265A 2001, Smith, et al. J. Mass Spectrometry 32:135-146 1997), a number of improvements to their methodologies and experimental equipment have been developed and implemented which have significantly improved throughput, comprehensiveness, and resolution. They term these collective enhancements high throughput-high resolution Deuterium Exchange-Mass Spectrometry (DXMS). As described herein, it is a methodology well suited to provide data to refine high throughput structure determination.

COREX: Development of Reliable Methods for Calculating Amide Hydrogen Exchange Rates from Protein Structures

[0008] Under native conditions, proteins are not uniform, individual structures, but actually ensembles of conformational states. This observation led to the development of the COREX algorithm, a computational tool that utilizes the high-resolution structure as a template to generate a large ensemble of incrementally different conformational states. COREX has been shown to predict exchange rates with remarkable accuracy and precision when tested against available NMR-derived experimental data, suggesting that the calculated

ensemble captures the general features of the actual ensemble, and thus provides a realistic physical description of proteins. The COREX algorithm is a structure-based thermodynamic model in which proteins are represented as ensembles of conformations rather than as discrete structures (Hilser, et al. J Mol Biol 262:756-72. 1996, Hilser, et al. Proteins 26:123-33. 1996, Hilser, et al. Biophys Chem 64:69-79. 1997, Hilser, et al. Proteins 27:171-83. 1997, Hilser, et al. Proc Natl Acad Sci U S A 95:9903-8. 1998), reviewed in (Hilser Methods Mol Biol 168:93-116. 2001). This algorithm performs the following two computational tasks (I) The high resolution X-ray or NMR structure (or a presumed structure) is used as a template from which a large ensemble of conformations ($>10^5$) is generated. (II) The relative enthalpy, ΔH_i, and entropy, ΔS_i, are calculated for each conformation using a surface area-based parameterization of the energetics (Murphy, et al. J Mol Biol 227:293-306. 1992, Murphy, et al. Adv Protein Chem 43:313-61. 1992, Gomez, et al. Proteins 22:404-12. 1995, Habermann, et al. Protein Sci 5:1229-39. 1996, Xie, et al. Protein Sci 3:2175-84. 1994, Lee, et al. Proteins 20:68-84. 1994, D'Aquino, et al. Proteins 25:143-56. 1996, Baldwin Proc Natl Acad Sci U S A 83:8069-72. 1986, Luque, et al. Biochemistry 35:13681-8. 1996) and the resultant Gibbs energy change, ΔG_i, is used to calculate the probability of each state.

[0009] Ensemble-Based Environmental Descriptors. The most important aspect of the ensemble-based approach is the ability to calculate the probability for each residue to be in a folded or an unfolded conformation (Hilser, et al. J Mol Biol 262:756-72. 1996, Hilser, et al. Proteins 26:123-33. 1996, Hilser, et al. Biophys Chem 64:69-79. 1997, Hilser, et al. Proteins 27:171-83. 1997). These residue-specific probabilities are used to define residue stability constants, $K_{f,j}$, as:

$$\kappa_{f,j} = \frac{\Sigma\ P_{f,j}}{\Sigma\ P_{nf,j}} \qquad (1)$$

where $\Sigma P_{f,j}$ and $\Sigma P_{nf,j}$ are the summed probabilities of all states in the ensemble in which residue j is either folded or unfolded, respectively. According to equation (1), residues with high stability constants will be folded in the majority of highly probable states, while residues with low constants will be unfolded in those states.

[0010] The significance of equation 1 is two- fold. First, the values can be compared to protection factors obtained from hydrogen exchange measurements, and thus can be verified experimentally. To this end, within COREX an ability to derive exchange rate values from the ensemble behavior was implemented, and demonstrated that such COREX- predicted rates were remarkably accurate when compared with existing NMR-determined rate data (Hilser, et al. J Mol Biol 262:756-72. 1996, Hilser, et al. Proteins 26:123-33. 1996, Hilser, et al. Biophys Chem 64:69-79. 1997, Hilser, et al. Proteins 27:171-83. 1997, Hilser, et al. Proc Natl Acad Sci U S A 95:9903-8. 1998). Thus the fundamental ability of COREX to calculate reasonably accurate exchange rate fingerprints has been established. It is important to note that COREX was not developed with the aim of predicting hydrogen exchange rates *per se*, and straightforward refinements to the method of calculating amide hydrogen exchange rates presently within COREX are likely to significantly improve the accuracy of COREX-predicted exchange rates. The availability of high resolution, comprehensive experimental measurements of exchange rates through DXMS will allow validation of such algorithmic and empiric adjustments. Considerable evidence indicates that measurements of amide hydrogen exchange rates with mass spectrometry match and extend those from NMR (Zhang, et al. Protein Sci 6:2203-17. 1997, Chung, et al. Protein Sci 6:1316-24. 1997).

[0011] Second, the residue-specific stability constants in equation (1) can be used as structural-energetic descriptors of the environment for each residue. As the energy of any state is derived from the structure, and includes terms for surface exposure, conformational entropy, *etc.*, the stability constant for each residue provides an implicit weighting of the different parameters which describe the residue environment. This provides the opportunity for residue-specific thermodynamic measurements (for example provided by DXMS-measured exchange rates) to provide constraints to each residue's environment, and likely dramatically improve success rates for fold prediction, a structure determination approach that is now described.

[0012] COREX-Based Prediction of Protein Structures. One of the most important problems in the area of protein folding is to identify what unique fold a particular sequence of amino acids will adopt. Two approaches that have become popular are comparative (*i.e.*, homology) modeling and fold recognition. While they differ in many regards, both strategies

use physico-chemical information about amino acid properties in known structures to derive the most probable model for an unknown sequence. A unique strategy for addressing fold recognition and prediction has been developed that is based on a previously developed statistical description of proteins that effectively links functional information with sequence information through the high-resolution structure. This strategy uses the functional information (i.e. the ensemble of states and their relative probabilities) derived from the COREX algorithm by mapping it back onto sequence space to identify fold or functional relationships between sequences. (Wrabl, J.O., S.A. Larson, and V.J. Hilser, Thermodynamic environments in proteins: Fundamental determinants of fold specificity. Protein Sci, 2002. 11(8): p. 1945-57., Hilser, V.J. and E. Freire, Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. J Mol Biol, 1996. 262(5): p. 756-72., Hilser, V.J., J. Gomez, and E. Freire, The enthalpy change in protein folding and binding: refinement of parameters for structure-based calculations. Proteins, 1996. 26(2): p. 123-33.. Hilser, V.J., B.D. Townsend, and E. Freire, Structure-based statistical thermodynamic analysis of T4 lysozyme mutants: structural mapping of cooperative interactions. Biophys Chem, 1997. 64(1-3): p. 69-79., Hilser, V.J. and E. Freire, Predicting the equilibrium protein folding pathway: structure-based analysis of staphylococcal nuclease. Proteins, 1997. 27(2): p. 171-83., Hilser, V.J., D. Dowdy, T.G. Oas, and E. Freire, The structural distribution of cooperative interactions in proteins: analysis of the native state ensemble. Proc Natl Acad Sci U S A, 1998. 95(17): p. 9903-8.)

[0013] Examples of Fold-Specific stability fingerprints can generated by COREX. Analyses have been performed on multiple members of different fold classes and "fold-specific" libraries for several structural motifs including protein SH3 domains have been compiled. In spite of any differences, the regional variations in stability constants seen are remarkably similar for the different SH3 domains. This result, which has also been found for other classes of proteins, suggests that different folds have structural thermodynamic signatures that are more or less invariant with sequence.

[0014] Thermodynamic Environments in Proteins. Stability constants provide a residue-specific description of the regional differences in stability within a protein structure. The importance of this quantity from the point of view of fold recognition is two-fold. First the

stability constant can be compared directly to protection factors obtained from native state hydrogen exchange experiments, thus providing an experimentally verifiable residue-specific description of the ensemble. Second, as amino acids are non-randomly distributed across high, medium and low stability environments, the stability constant as a function of residue position provides a convenient 1-dimensional representation of the 3-dimensional structure. It has been established that such a description contains significant structure-encoding information (Wrabl, et al. Protein Sci 10:1032-45. 2001). Although the stability constants provide a residue specific description of the stability in various regions of the protein, the origins of the stability may differ for each region. For instance, in SH3 domains the RT loop (residues 15-25) and the distal loop (residues 45-55) each have low stability constants.

[0015] As the free energy of each state generated by the COREX algorithm is determined from apolar, polar, and conformational entropy contributions:

$$\Delta G_i = \Delta G_{apolar,i} + \Delta G_{polar,i} + \Delta G_{confS,i}$$

the energetics of each residue in a protein can be characterized by a minimum of three energetic descriptors as described (Wrabl, J.O., S.A. Larson, and V.J. Hilser, Thermodynamic environments in proteins: Fundamental determinants of fold specificity. Protein Sci, 2002. 11(8): p. 1945-57) Together these descriptors can determine the enthalpic and entropic contributions to stability. how much of the solvation contribution is due to apolar and polar surface, and of the total stability, what fraction is due to conformational entropy as opposed to solvation.

Rosetta

[0016] The Rosetta method of *de novo* protein-structure prediction is based on the assumption that the distribution of conformations available to any short segment of the chain is determined largely by the local sequence. To approximate the conformational space available to each segment, sets of 3-mer and 9-mer fragments for each position along the chain are extracted from the protein-structure database based on the sequence-profile similarity and secondary-structure predictions. Compact structures are then assembled by randomly combining these fragments using a Monte Carlo simulated annealing search. The fitness of individual conformations with respect to non-local interactions is evaluated using an energy function derived from observed distributions in known protein structures. The

energy function favors hydrophobic burial and strand pairing, and disfavors steric clashes. For each target sequence, large numbers (1,000 - 10,000) of possible structures (termed "decoys" in the Rosetta literature) are generated with this protocol.

[0017] As presently employed, the population of decoys is automatically filtered and then refined in a full-atom protocol that adds on all side-chain and hydrogen atoms and performs a coupled Monte Carlo minimization of the backbone and side-chain conformations. In addition to the energy function described above, the full-atom energy function includes Lennard-Jones and pairwise solvation potentials, as well as several statistical potentials for side-chain atom pairs, side-chain rotamers, and hydrogen bonds. The accuracy of the Rosetta full-atom energy function has been demonstrated recently by the experimental verification of a computationally designed novel fold (Kuhlman et al. 2003).

[0018] Despite the relative success of the Rosetta method, there is a pressing need for methods to generate "better", more accurate Rosetta predictions. For the purpose of generating more accurate structures with Rosetta, some of the most useful and rapidly attainable information from DXMS is the identification of amide hydrogens that exchange at the maximal rate. These "fast-amides" are hydrogen-bonded to water rather than the protein most of the time, so "fast-amides" might only participate in loops, in kinks in α-helices, and in edge strands of β-sheets. Indeed, contiguous stretches of four or more "fast-amides" identified by DXMS have been found to map to disordered regions of proteins (see, for example, Examples herein), and the same work reveals the identity and location of the isolated (single, pairs, triplets) of very rapidly exchanging amides that are within structured regions of proteins, thereby revealing that they are disposed on the surface of the protein.

[0019] Knowledge of the identity and locations of these surface disposed, very fast-exchanging amides, can readily be incorporated into any computational structure prediction method, thereby greatly refining prediction accuracy, and decreasing needed computing time. For example in the Rosetta algorithm, the "very fast-amide" locations can be incorporated into the fragment-picking step, by requiring that all fragments spanning a fast-amide are consistent with that amide lacking a hydrogen bond. Though most structures in the database do not contain hydrogens, Rosetta is capable of placing backbone hydrogens in a structure in a highly accurate manner. Second, the "fast-amide" locations can be incorporated into the

folding protocol by adding a term to the scoring function that favors exposure (fewer neighbors) for these locations. Furthermore, the "fast-amide" locations can be incorporated into the full-atom refinement protocol both by requiring that those amides do not form hydrogen bonds and by requiring that those amides have significant solvent-exposed-surface-area. Finally these constraints can be applied by modifying the polypeptide covalent sequence that is used in the folding protocol is such a way that the amide of each identified very fast exchanger is modified to include a covalent structure which extends from the amide as a cone, being equivalent in three dimensional space to a cluster of hydrogen- bonded water molecules (preferably 5-10 molecules) extending out from the amide. In the folding protocol, each of these "very fast exchanging amide-decorated cones" is given approximately the same restrictions on space violations as other parts of the polypeptide sequence (atoms that are not covalently bonded are not allowed to "fall" inside each other) except that the "water cone" decorations are allowed to violate each other freely. In this manner the folding protocol will reserve an unimpeded route for water molecules to freely approach and interact with the "very fast amides" in any resulting structure. These additions to Rosetta should improve the prediction of edge beta strands and hence should be highly useful for structure prediction in all-beta proteins that have proven particularly challenging for Rosetta.

[0020] The incorporation of the "fast-amide" data into the Rosetta structure prediction can be refined and validated using standard techniques. A training set of ~30 proteins with known structures and diverse scop classes and sizes can be established, and the "fast-amide" locations determined by DXMS for all of these proteins. Rosetta decoys are then generated for all proteins in the training set, both with and without the incorporation of the fast-amide data, and improvements in the accuracy of the predictions assessed by comparing rms-to-native distributions of the decoys generated with and without the fast-amide data. This type of comparison can be made each time a significant change is made in the incorporation of the fast-amide data into Rosetta.

[0021] Advanced methods for amide hydrogen exchange rate fingerprint comparison. In calculating the goodness of fit between the experimental hydrogen exchange fingerprint and the COREX-produced fingerprints, two measures have considered, namely mean absolute error (mae) and root mean squared error (rmse). Examining the patterns of individual amide

errors between the experimental rate and the COREX rate for any given candidate structure, it is evident that extreme outliers are a common feature; that is, there are typically at least one and often many individual amides for which the difference between the experimental and COREX rates is quite large relative to the differences for other amides. Mean absolute error and root mean squared error are ideally suited to accommodating random variation in errors arising from the double exponential and normal distributions respectively, but can be seriously adversely affected by outliers. The observed outliers in the fingerprint matching analysis are far from what one would expect from either the double exponential or normal distributions. It follows that in the presence of such outliers, using mae or rmse to draw inferences about which COREX structure best fits the experimental data presents the serious danger of discarding a correct (or nearly correct) structure for which experimental and COREX results are mostly in agreement but suffer from large differences at a modest number of amides and instead selecting an incorrect structure that agrees less well overall with experiment, but whose errors (especially the outliers) are more moderate.

[0022] The adverse effects of outliers have been well studied in the statistics literature, with particular attention to measuring goodness of fit in the presence of outliers (Sakata and White, 1995; Sakata and White, 1998). To specify the robust goodness of fit measures it is possible to define the amide-specific error $u_i = x_i^* - x_i$, where x_i^* is the deconvoluted experimental hydrogen exchange rate for amide i, and x_i is the COREX-determined rate for amide i (for a given candidate structure). A general family of measures of goodness of fit (also known as "scale estimators") is given by

$$z(r, K) = \inf \{ s > 0 : m^{-1} S_i \; r(u_i / s) < K \},$$

where r is a given continuous even function satisfying $r(0) = 0$, together with further technical conditions (see Sakata and White, 1998, p. 555), K is a given constant, and m is the total number of amides. The family is indexed by r and K; different choices for r and K give goodness of fit measures with differing properties. For example, setting $K = \infty$ and taking $r(w) = w^2$ in z gives the mean squared error measure, whereas taking $r(w) = |w|$ gives the mean absolute error measure. Choosing r to be non-decreasing and bounded endows $z(r, K)$ with appealing robustness ("high breakdown") properties against outliers. In particular, the choice

$$r(w) = w^2/2 - w^4 / 2d^2 + w^6/ 6\,d^4 \quad \text{for } | w | \leq d$$

$$r(w) = d^2/6 \quad \text{for } | w | > d,$$

where d is a given constant, is a choice derived from Tukey's biweight function (see Beaton and Tukey, 1975, and Rousseeuw and Yohai, 1984) that can provide z(r, K) with substantial protection against the adverse effects of outliers. In a preferred embodiment, use is made of this biweight-based goodness of fit measure and other similar measures to assess and rank the quality of the matches between the experimental fingerprint and those produced by COREX, thereby avioding the problems created by outliers and increasing the ability to determine promising structures. Note that when using the biweight-based measure r there are still two parameters to be chosen: d and K. Minimal experimentation will be required to establish regions of (d, K) values where the goodness-of-fit ranking of the available candidate structures is relatively stable. (Beaton, A. and J.W. Tukey (1975), "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data," *Technometrics* 16, 147-192., Rousseeuw, P. and V. Yohai (1984), "Robust Regression by means of *S*-Estimators," in *Robust and Nonlinear Time Series Analysis*, ed. W.H. Franke and D. Martin,. New York: Springer-Verlag, pp. 256-272., Sakata S. and H. White (1995), "An Alternative Definition of Finite Sample Breakdown Point With Applications to Regression Model Estimators," *Journal of the American Statistical Association* 90, 1099-1106., Sakata, S. and H. White (1998), "High Breakdown Point Conditional Dispersion Estimation with Application to S&P 500 Daily Returns Volatility," *Econometrica* 66, 529-567)., White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica* 50, 1-25).

Hydrogen (Proton) Exchange

[0023] When a protein in its native folded state is incubated in buffers containing an isotope of hydrogen (for example, tritium or deuterium labeled water), isotope in the buffer reversibly exchanges with normal hydrogen present in the protein at acidic positions (for example, -OH, -SH, and -NH groups) with rates of exchange which are dependent on each exchangeable hydrogen's chemical environment, temperature, and most importantly, its accessibility to the isotope of hydrogen present in the buffer (see, *e.g.*, Englander *et al.*, *Meth. Enzymol.* **49**:24-39, 1978; Englander *et al., Meth. Enzymol.* **26**:406-413, 1972). Accessibility is determined in turn by both the surface (solvent-exposed) disposition of the hydrogen, and

the degree to which it is hydrogen-bonded to other regions of the folded polypeptide. Simply stated, an acidic hydrogen present on amino acid residues which are on the outside (buffer-exposed) surface of the protein and which are hydrogen-bonded to solvent water will often exchange more rapidly with heavy hydrogen in the buffer than will a similar acidic hydrogen which is buried and hydrogen-bonded within the folded polypeptide. Hydrogen exchange reactions can be greatly accelerated by both acid and base-mediated catalysis; and the rate of exchange observed at any particular pH is the sum of both acid and base mediated mechanisms. For many acidic hydrogens, a pH of 2.2 - 2.7 results in an overall minimum rate of exchange (Englander *et al., Anal. Biochem.* **147**:234-244, 1985; Englander *et al., Biopolymers* 7:379-393, 1969; Molday *et al., Biochemistry* **11**:150, 1972; Kim *et al., Biochemistry* **21**:1, 1982; Bai *et al., Proteins: Struct. Funct. Genet.* **17**:75-86, 1993; and Connelly *et al., Proteins: Struct. Funct. Genet.* **17**:87-92). While hydrogens in protein hydroxyl and amino groups exchange with tritium or deuterium in buffer at millisecond rates, the exchange rate of one particular acidic hydrogen, the peptide amide bond hydrogen, is considerably slower, having a half life of exchange (when freely accessible, and freely hydrogen-bonded to solvent water) of approximately 0.5 seconds at 0 °C, pH 7, which is greatly slowed to a half life of exchange of 70 minutes at 0 °C, pH 2.7. When a polypeptide is in a denatured, unstructured configuration (also termed a "random coil") all of its amide hydrogens can freely exchange with solvent hydrogen. However, the precise rate of exchange varies up to 200 fold from amide to amide in such unstructured configurations, the rate of exchange at each particular amide being determined by localized primary amino acid sequence-dependent effects that can be calculated from a knowledge of the peptide's primary sequence (Bai *et al., supra*). When peptide amide hydrogens are buried within a folded polypeptide, or are hydrogen bonded to other parts of the polypeptide, exchange half-lives with solvent hydrogens are often considerably lengthened, at times being measured in hours to days.

[0024] Hydrogen exchange at peptide amides is a fully reversible reaction, and rates of on-exchange (solvent deuterium replacing protein-bound normal hydrogen) are identical to rates of off-exchange (hydrogen replacing protein-bound deuterium) if the state of a particular peptide amide within a protein, including its chemical environment and accessibility to solvent hydrogens, remains identical during hydrogen exchange conditions.

[0025] Hydrogen exchange is commonly measured by performing studies with proteins and aqueous buffers that are differentially tagged with pairs of the three isotopic forms of hydrogen (^{1}H, normal hydrogen; ^{2}H, deuterium; ^{3}H, tritium). If the pair of normal hydrogen and tritium are employed, it is referred to as tritium exchange; if normal hydrogen and deuterium are employed, as deuterium exchange. Different physicochemical techniques are in general used to follow the distribution of the two isotopes in deuterium versus tritium exchange. The rates of exchange of other acidic protons (-OH, -NH, and -SH) are so rapid that they cannot be followed in these techniques and all subsequent discussion refers exclusively to peptide amide proton exchange.

Tritium Exchange Techniques

[0026] Tritium exchange techniques (where the amount of the isotope is determined by radioactivity measurements) have been extensively used for the measurement of peptide amide exchange rates within an individual protein. In these studies, purified proteins are on-exchanged by incubation in buffers containing tritiated water for varying periods of time, optionally transferred to buffers free of tritium, and the rate of off-exchange of tritium determined. By analysis of the rates of tritium on-and off-exchange, estimates of the numbers of peptide amide protons in the protein whose exchange rates fall within particular exchange rate ranges can be made. These studies do not allow a determination of the identity (location within the protein's primary amino acid sequence) of the exchanging amide hydrogens measured.

[0027] Extensions of these techniques have been used to detect the presence within proteins of peptide amides which experience allosterically-induced changes in their local chemical environment and to study pathways of protein folding (Englander *et al., Meth. Enzymol.* **26**:406-413, 1972; Englander *et al., J. Biol. Chem.* **248**:4852-4861, 1973; Englander, *Biochemistry* **26**:1846-1850, 1987; Louie *et al., J. Mol. Biol.* **201**:765-772, 1988). For these studies, tritium on-exchanged proteins are often allowed to off-exchange after they have experienced either an allosteric change, or have undergone time-dependent folding upon themselves, and the number of peptide amide hydrogens which experience a change in their exchange rate subsequent to the allosteric/folding modifications determined. Changes in exchange rate indicate that alterations of the chemical environment of particular peptide amides have occurred which are relevant to proton exchange (solvent accessibility, hydrogen

bonding, *etc.*). Peptide amide hydrogens which undergo an induced slowing in their exchange rate are referred to as "slowed amides" and if previously on-exchanged tritium is sufficiently slowed in its off-exchange from such amides there results a "functional tritium labeling" of these amides. From these measurements, inferences are made as to the structural nature of the shape changes which occurred within the isolated protein. Again, determination of the identity of the particular peptide amides experiencing changes in their environment is not possible with these techniques.

[0028] Several investigators have described technical extensions (collectively referred to as "medium resolution tritium exchange") which allow the locations of particular slowed, tritium labeled peptide amides within the primary sequence of small proteins to be localized to a particular proteolytic fragment, though not to a particular amino acid.

[0029] Rosa and Richards were the first to describe and utilize medium resolution tritium techniques in their studies of the folding of ribonuclease S protein fragments (Rosa *et al., J. Mol. Biol.* **133**:399-416, 1979; Rosa *et al., J. Mol. Biol.* **145**:835-851, 1981; and Rosa *et al., J. Mol. Biol.* **160**:517-530, 1982). However, the techniques described by Rosa and Richards were of marginal utility, primarily due to their failure to optimize certain critical experimental steps. No studies employing related techniques were published until the work of Englander and co-workers in which extensive modifications and optimizations of the Rosa and Richards technique were first described.

[0030] Englander's investigations utilizing tritium exchange have focused exclusively on the study of allosteric changes which take place in tetrameric hemoglobin (a subunit and b subunit 16 kD in size each) upon deoxygenation (Englander *et al., Biophys. J.* **10**:577, 1979; Rogero *et al., Meth. Enzymol.* **131**:508-517, 1986; Ray *et al., Biochemistry* **25**:3000-3007, 1986; and Louie *et al., J. Mol. Biol.* **201**:755-764, 1988). In the Englander procedure, native hemoglobin in the oxygenated state is on-exchanged in tritiated water. The hemoglobin is then deoxygenated (inducing allosteric change), transferred to tritium-free buffers by gel permeation column chromatography, and then allowed to off-exchange for 10 - 50 times the on-exchange time. On-exchanged tritium present on peptide amides which experience no change in exchange rate subsequent to the induced allosteric change in hemoglobin structure off-exchanges at rates identical to its on-exchange rates, and therefore is almost totally removed from the protein after the long off-exchange period. However, peptide amides

which experience slowing of their exchange rate subsequent to the induced allosteric changes preferentially retain the tritium label during the period of off-exchange.

[0031] To localize (in terms of hemoglobin's primary sequence) the slowed amides bearing the residual tritium label, Englander then proteolytically fragments the off-exchanged hemoglobin with the protease pepsin, separates, isolates and identifies the various peptide fragments by reverse phase high pressure liquid chromatography (RP-HPLC), and determines which fragments bear the residual tritium label by scintillation counting. However, as the fragmentation of hemoglobin proceeds, each fragment's secondary and tertiary structure is lost and the unfolded peptide amide hydrogens become freely accessible to H_2O in the buffer. At physiologic pH (>6), any amide-bound tritium label would leave the unfolded fragments within seconds. Englander therefore performs the fragmentation and HPLC peptide isolation procedures under conditions which minimize peptide amide proton exchange, including cold temperature (4 °C) and use of phosphate buffers at pH 2.7. This technique has been used successfully by Englander to coarsely identify and localize the peptide regions of hemoglobin α and β chains which participate in deoxygenation-induced allosteric changes. The ability of the Englander technique to localize tritium labeled amides, while an important advance, remains low; at best, Englander reports that his technique localizes amide tritium label to hemoglobin peptides 14 amino acids or greater in size, without the ability to further sublocalize the label. Moreover, in Englander's work, there is no appreciation that a suitably adapted exchange technique might be used to identify the peptide amides which reside in the contacting surface of a protein receptor and its binding partner. Instead, these Englander disclosures are concerned with the mapping of allosteric changes in hemoglobin.

[0032] Unfortunately, acid proteases are very nonspecific in their sites of cleavage, leading to considerable HPLC separation difficulties. Englander tried to work around these problems, for the localization of hemoglobin peptides experiencing allosteric changes, by taking advantage of the fact that some peptide bonds are somewhat more sensitive to pepsin than others. Even then, the fragments were "difficult to separate cleanly". They were also, of course, longer (on average), and therefore the resolution was lower. Englander concludes, "At present the total analysis of the HX (hydrogen exchange) behavior of a given protein by these methods is an immense task. In a large sense, the best strategies for undertaking such a task remain to be formulated. Also, these efforts would benefit from further technical

improvements, for example in HPLC separation capability and perhaps especially in the development of additional acid proteases with properties adapted to the needs of these experiments" (Englander *et al., Anal. Biochem.* **147**:234-244, 1985).

[0033] Over the succeeding years since this observation was made, no advances have been disclosed which address these critical limitations of the medium resolution hydrogen exchange technique. Most acid-reactive proteases are in general no more specific in their cleavage patterns than pepsin. Efforts to improve the technology by employing other acid reactive proteases other than pepsin have not significantly improved the technique. Allewell and co-workers have disclosed studies utilizing the Englander techniques to localize induced allosteric changes in the enzyme *Escherichia coli* aspartate transcarbamylase (Burz *et al., Biophys. J.* **49**:70-72, 1986; Mallikarachchi *et al., Biochemistry* **28**:5386-5391, 1989). Burz *et al.* is a brief disclosure in which the isolated R2 subunit of this enzyme is on-exchanged in tritiated buffer of specific activity 100 mCi/ml, allosteric change induced by the addition of ATP, and then the conformationally altered subunit off-exchanged. The enzyme R2 subunit was then proteolytically cleaved with pepsin and analyzed for the amount of label present in certain fragments. Analysis employed techniques which rigidly adhered to the recommendations of Englander, utilizing a single RP-HPLC separation in a pH 2.8 buffer.

[0034] ATP binding to the enzyme was shown to alter the rate of exchange of hydrogens within several relatively large peptide fragments of the R2 subunit. In a subsequent more complete disclosure (Mallikarachchi, *supra*), the Allewell group discloses studies of the allosteric changes induced in the R2 subunit by both ATP and CTP. They disclose on-exchange of the R2 subunit in tritiated water-containing buffer of specific activity 22-45 mCi/ml, addition of ATP or CTP followed by off-exchange of the tritium in normal water-containing buffer. The analysis comprised digestion of the complex with pepsin, and separation of the peptide fragments by reverse phase HPLC in a pH 2.8 or pH 2.7 buffer, all of which rigidly adheres to the teachings of Englander. Peptides were identified by amino acid composition or by N-terminal analysis, and the radioactivity of each fragment was determined by scintillation counting. In both of these studies the localization of tritium label was limited to peptides which averaged 10-15 amino acids in size, without higher resolution being attempted. Beasty *et al.*, (*Biochemistry* **24**:3547-3553, 1985) have disclosed studies employing tritium exchange techniques to study folding of the α subunit of E. coli tryptophan

synthetase. The authors employed tritiated water of specific activity 20 mCi/ml, and fragmented the tritium labeled enzyme protein with trypsin at a pH 5.5, conditions under which the protein and the large fragments generated retained sufficient folded structure to protect amide hydrogens from off-exchange during proteolysis and HPLC analysis. Under these conditions, the authors were able to produce only 3 protein fragments, the smallest being 70 amino acids in size. The authors made no further attempt to sublocalize the label by further digestion and/or HPLC analysis. Indeed, under the experimental conditions they employed (they performed all steps at 12 °C instead of 4 °C, and performed proteolysis at pH 5.5 instead of pH in the range of 2-3), it would have been impossible to further sublocalize the labeled amides by tritium exchange, as label would have been immediately lost (off-exchanged) by the unfolding of subsequently generated proteolytic fragments at pH 5.5 if they were less than 10-30 amino acids in size. Additional references disclosing tritium exchange methods include Fromageot *et al.*, U.S. Patent No. 3,828,102, which discloses using hydrogen exchange to tritium label a protein and its binding partner, and Benson, U.S. Pat. Nos. 3,560,158 and 3,623,840, which discloses using hydrogen exchange to tritiate compounds for analytical purposes.

Deuterium Exchange Techniques

[0035] Fesik *et al.* (*Biochem. Biophys. Res. Commun.* **147**:892-898, 1987) disclose measuring by NMR the hydrogen (deuterium) exchange of a peptide before and after it is bound to a protein. From this data, the interactions of various hydrogens in the peptide with the binding site of the protein are analyzed. Paterson *et al.* (*Science* **249**:755-759, 1990) and Mayne *et al.* (*Biochemistry* **31**:10678-10685, 1992) disclose NMR mapping of an antibody binding site on a protein (cytochrome-C) using deuterium exchange. This relatively small protein, with a solved NMR structure, is first complexed to anti-cytochrome-C monoclonal antibody, and the preformed complex then incubated in deuterated water-containing buffers and NMR spectra obtained at several time intervals. The NMR spectrum of the antigen-antibody complex is examined for the peptide amides which experience slowed hydrogen exchange with solvent deuterium as compared to their rate of exchange in uncomplexed native cytochrome-C. Benjamin *et al.* (*Biochemistry* **31**:9539-0545, 1992) employ an identical NMR-deuterium technique to study the interaction of hen egg lysozyme (HEL) with HEL-specific monoclonal antibodies. While both this NMR-deuterium technique, and

medium resolution tritium exchange rely on the phenomenon of proton exchange at peptide amides, they utilize radically different methodologies to measure and localize the exchanging amide hydrogens. Furthermore, study of proteins by the NMR technique is not possible unless the protein is small (generally less than 30 kD), large amounts of the protein are available for the study, and computationally intensive resonance assignment work is completed.

[0036] Subsequently, others have disclosed techniques in which exchange-deuterated proteins are incubated with binding partner, off-exchanged, the complex fragmented with pepsin, and deuterium-bearing peptides identified by single stage fast atom bombardment (Fab) or electrospray mass spectroscopy (MS) (Thevenon-Emeric *et al., Anal. Chem.* **64**:2456-2358, 1992; Winger *et al., J. Am. Chem. Soc.* **114**:5897-5989, 1992; Zhang *et al., Prot. Sci.* **2**:522-531, 1993; Katta *et al., J. Am. Chem. Soc.* **115**:6317-6321, 1993; and Chi *et al., Org. Mass Spectrometry* 7:58-62, 1993; Engen and Smith, *Anal. Chem.* **73**:256A- 265A, 2001; Englander *et al., Protein Sci.* **6**: 1101-1109, 1997; Dharmasiri and Smith, *Anal. Chem.* **68**:2340-2344, 1996; Smith *et al., J. Mass Spectrometry* **32**:135-146, 1997; Deng and Smith, *Biochemistry* **37**:6256-6262, 1998). In these studies, only the enzyme pepsin is employed to effect enzymatic fragmentation under slowed exchange conditions, and no attempt made to increase the number and quantity of useful fragments produced and studied beyond employing the methods disclosed by Englander and colleagues some decades prior. The resolution of the deuterium-exchange mass spectrometry work disclosed in these publications therefore remained at the 10-14 amino acid level, with the primary limitation of their art being the ability to generate only a small number of peptides with the endopeptidase pepsin, as they employed it.

[0037] U.S. Patent Nos. 5,658,739; 6,291,189; and 6,331,400 issued to Woods, Jr. (each of which is hereby incorporated by reference herein in its entirety), disclose improved methods of determining polypeptide structure and binding sites utilizing hydrogen-exchange-labeled peptide amides, importantly including a method of increasing the resolution of the technique to the 1-5 amino acid level. This increased ability to more precisely localize exchanged amide hydrogens was afforded by the novel use of acid-resistant carboxypeptidases to effect a subsequent progressive sub-fragmentation of the small number of relatively large-sized pepsin-generated peptides initially produced in the method. In these

prior methods, finer localization of the labels is achieved by analysis of subfragments generated by controlled, stepwise, sub-degradation ("progressive degradation") of each pepsin-generated, labeled peptide under slowed exchange conditions. According to these prior methods, the protein or a peptide fragment is said to be "progressively", "stepwise" or "sequentially" degraded if a series of fragments are obtained which are similar to those which would be achieved with an ideal exopeptidase. Carboxypeptidase-P, carboxypeptidase Y, and several other acid-reactive (*i.e.*, enzymatically active under acid conditions) carboxypeptidases are specified for use in said progressive degradation of peptides under acidic conditions. To date, no aminopeptidases have been reported that are acid resistant; as a practicality, the only exopeptidases known or likely to be useful for this method are therefore carboxypeptidases.

[0038] By performing such measurement of the exchange rates of peptide amide hydrogens within a protein, one can determine its stability at the individual amino acid level. Ranking and comparison of the exchange rates of a protein's amide hydrogens therefore allows direct identification and localization of structured versus unstructured regions of the protein. Despite the utility of such exchange data, the methods used to obtain it have remained labor intensive and time consuming, with substantial limitations in throughput, comprehensiveness and resolution.

SUMMARY OF THE INVENTION

[0039] The present invention provides methods for determining polypeptide and protein three-dimensional structures. In a particular aspect, the invention relates to methods for three-dimensional structure determination that employ hydrogen exchange analysis to refine, constrain and improve computational protein structure predictive methods.

[0040] Preferred methods of the present invention employ novel high resolution hydrogen exchange analysis. In some embodiments of the invention, methods of hydrogen exchange analysis comprise fragmentation of a labeled protein using methods described in U.S. Patent Nos. 5,658,739; 6,331,400, and 6,291,189, the entire disclosures of which are incorporated herein by reference. In other embodiments of the invention, the hydrogen exchange analysis

allows for high-throughput structural determinations due to simplifications of the protein fragmentation methods described in U.S. Patent Nos. 5,658,739; 6,331,400, and 6,291,189.

[0041] According to a first aspect of the present invention, there are provided methods of structure prediction and/or determination of a protein of interest of unknown structure. These methods comprise comparing calculated rates of amide hydrogen exchange determined for a set of predicted possible structures for said protein of interest with experimental hydrogen exchange analysis of said protein of interest, and identifying one or more structures from said set of predicted possible structures having a calculated exchange rate profile closely matching the experimental exchange rate profile.

[0042] In general, the protein may be studied by mass spectrometry based hydrogen exchange methods, or NMR methods to measure amide hydrogen exchange rates, to establish the protein's true amide hydrogen exchange profile, or exchange rate fingerprint. A simple analysis of a portion of this rate information allows precise identification of the protein's peptide amides (typically 10-20% of them) that have very fast exchange rates, indicating that they are always in full contact with solvent water in the protein, and therefore are on its surface. Multiple structures (preferably 1,000-10,000) may be predicted/ proposed for the target protein using any of a number of structure- predicting methods, including the Rosetta algorithm, with the computations performed in a manner that takes advantage of the foregoing derived knowledge of the identity of the surface-disposed amides, greatly improving the accuracy of predictions and speeding calculations. Methods capable of estimating or calculating the likely exchange rates of the amides in proposed or actual 3D structures, including the COREX algorithm, are used to construct virtual hydrogen-exchange rate fingerprints or profiles for each of the several proposed structure(s) for the target protein. These calculated fingerprints are compared to the true experimentally determined rate fingerprint by any of a number of methods for such comparisons, and the structural predictions with calculated exchange rate fingerprints most closely matching experimentally determined fingerprints identified.

[0043] The principal virtues of this approach are its simplicity, and the ease with which hydrogen exchange data can be rapidly obtained despite the idiosyncrasies of the protein under study. Most of the experimental technique is performed under conditions that suppress

the unique features of individual amino acids - the use of acid pH, denaturants, and non-specific proteases, making the same basic hydrogen exchange methods universally applicable to proteins that have dramatically differing properties under native conditions.

[0044] In one embodiment, invention methods may be used to refine structure prediction for isolated, purified proteins. In various other embodiments, the invention methods may be used to refine structure prediction for complexes of proteins, or proteins bound to non-protein ligands.

[0045] In another embodiment, invention methods may be used to refine structure predictions for proteins that are under study by other means, including x-ray crystallography or NMR methods. Refined structure predictions provided by this method may provide model structures or templates that can facilitate the molecular replacement step of crystallographic protein structure determination. In molecular replacement, the structural coordinates of a structurally known protein thought to be homologous in structure to the unknown protein (typically based on primary sequence homology between structurally known and unknown protein) are used to generate a provisional model of the unknown protein by orienting and positioning the structural coordinates of the known protein within the unit cell of the unknown crystal so as best to account for the observed diffraction pattern of the unknown crystal, thereby facilitating phase determination (see, for example, paragraph [0115]). In this embodiment, invention methods are used to produce predicted structure(s) for the unknown protein that is consistent and compatible with experimentally determined hydrogen exchange measurements made on the unknown protein. This hydrogen-exchange-refined structural prediction(s) is then used to generate a provisional model of the unknown protein by orienting and positioning the structural coordinates of the known protein within the unit cell of the unknown crystal so as best to account for the observed diffraction pattern of the unknown crystal, thereby facilitating phase determination and structure as described in greater detail, for example, in paragraph [0115] below.

[0046] In another embodiment, the ability to define the surface-disposed amides of a protein (very fast exchanging amides) is employed for structure refinement efforts without the use of the "DXMS-COREX" filter element.

[0047] In another embodiment, the hydrogen exchange information that is compared to determine each structure prediction's accuracy includes (i) experimental rate fingerprint measurements, derived from raw experimental DXMS deuterated fragment data that is deconvoluted to amide- specific rates; and (ii) "virtual" amide specific rates calculated (for example by COREX) from a predictions 3-D structure. This method makes use of manual or computational approaches (described herein) for the deconvolution of aggregate DXMS experimental data to amide- specific exchange rates.

[0048] In another embodiment, the hydrogen exchange information that is compared to determine each structure prediction's accuracy includes (i) raw experimental DXMS deuterated fragment data; and (ii) "virtual" raw experimental DXMS deuterated fragment data that is generated by first calculating the amide specific rates (for example by COREX) from a predictions 3-D structure, and, then, with knowledge of the on and off exchange times used to generate the DXMS-derived experimental data and knowledge of the experimental data's fragment identities, calculating the deuteration magnitude of each fragment, for each on and off time used in the generation of the experimental data. This approach does not require an experimental data deconvolution step, and is likely to have the virtue of being more tolerant to errors and inaccuracies in the experimental data.

[0049] In another embodiment, the hydrogen exchange information that is used to determine structure prediction accuracy by either of the above approaches consists of experimental alkyl- hydrogen exchange data for a protein, and modified forms of rate calculating methods that allow calculation of alkyl-exchange rates in presumed or actual structures. Such modifications are readily accomplished by using the same solvent accessibility and exchange criteria as are used presently in such methods to calculate amide hydrogen exchange rates, but apply them to alkyl- hydrogen exchangeable positioned in the amino acids of a protein.

[0050] In another embodiment, the several components of the method (prediction generation, experimental data acquisition, exchange rate calculation (*i.e.*, COREX) are not only performed in the sequential manner suggested above, but in a manner in which there is contemporaneous, simultaneous performance of some or all of the several steps to promote computational economy.

[0051] The hydrogen exchange analysis comprises determining the quantity of isotopic hydrogen and/or the rate of exchange of hydrogen at a plurality of peptide amide hydrogens exchanged for isotopic hydrogen in a protein labeled with a hydrogen isotope other than ^{1}H, such as deuterium or tritium.

[0052] In one preferred embodiment, hereinafter referred to as "progressive proteolysis" (as defined in U.S. Patent No. 6,291,189, column 7, line 58 through column 8, line 33) the process of determining the quantity of isotopic hydrogen and/or the rate of exchange comprises: (a) fragmenting the labeled protein into a plurality of fragments under slowed hydrogen exchange conditions; (b) identifying which fragments of the plurality of fragments are labeled with isotopic hydrogen; (c) progressively degrading each fragment of the plurality of fragments to obtain a series of subfragments, wherein each subfragment of the series is composed of about 1-5 fewer amino acid residues than the preceding subfragment in the series from one end but with preservation of the other end of the subfragment series; (d) measuring an amount of isotopic hydrogen associated with each subfragment; and (e) correlating said amount of isotopic hydrogen associated with each subfragment with an amino acid sequence of the fragment from which said subfragment was generated, thereby determining the quantity of isotopic hydrogen and/or the rate of exchange of a plurality of peptide amide hydrogens exchanged for isotopic hydrogen in a protein labeled with a hydrogen isotope other than ^{1}H.

[0053] In one aspect of the invention, the step of progressively degrading comprises contacting the fragments with an acid resistant carboxypeptidase, for example, carboxypeptidase P, carboxypeptidase Y, carboxypeptidase W, carboxypeptidase C, or combinations of any two or more thereof.

[0054] In another preferred embodiment of the invention, hereinafter referred to as the "improved proteolysis" method, the process of determining the quantity of isotopic hydrogen and/or the rate of exchange comprises: (a) generating a population of sequence overlapping fragments of said labeled protein by treatment with at least one endopeptidase or combination of endopeptidases under conditions of slowed hydrogen exchange, and then (b) deconvoluting fragmentation data acquired from said population of sequence-overlapping endopeptidase-generated fragments. This improved method dramatically speeds and

modulates the sites and patterns of proteolysis by endopeptidases so as to produce highly varied and highly efficient fragmentation of the labeled protein in a single step, thereby avoiding the use of carboxypeptidases completely.

[0055] In one aspect, endopeptidase fragments are generated by cleaving said protein with at least one endopeptidase selected from the group consisting of a serine endopeptidase, a cysteine endopeptidase, an aspartic endopeptidase, a metalloendopeptidase, and a threonine endopeptidase. In a preferred method, endopeptidase fragments are generated by cleaving said protein with pepsin. Alternatively, endopeptidase fragments may be generated by cleaving said protein with newlase or Aspergillus protease XIII, or by more than one endopeptidase used in combination.

[0056] In preferred embodiments, invention methods measure the mass of peptide fragments, for example, utilizing mass spectrometry, to determine the presence or absence and/or quantity of an isotope of hydrogen on an endopeptidase fragment. Fragmentation data is deconvoluted by comparing the quantity and rate of exchange of isotope(s) on a plurality of sequence-overlapping endopeptidase-generated fragments with the quantity and rate of exchange of isotope(s) on at least one other endopeptidase fragment, wherein said quantities are corrected for back-exchange in an amino acid sequence-specific manner.

[0057] In another aspect, the present invention provides alternative methods of structure prediction and/or determination of a protein of interest of unknown structure. These methods comprise comparing calculated rates of amide hydrogen exchange determined for a set of predicted possible structures for said protein of interest using thermodynamic parameters of each amino acid residue in said protein of interest defined by hydrogen exchange analysis with experimental hydrogen exchange analysis of said protein, and identifying one or more structures from said set of predicted possible structures having a calculated exchange rate profile closely matching the experimental exchange rate profile.

[0058] In yet another aspect of the present invention, there are provided methods of performing molecular replacement, said methods comprising orienting and positioning the structural coordinates for the three-dimensional structure prediction(s) for a protein obtained by the above-described methods within the crystallographically-obtained unit cell of the

structurally unknown protein, so as best to account for the observed diffraction pattern of the structurally unknown protein crystal. In a presently preferred embodiment, accurate structural predictions are identified by the degree to which the orienting and positioning of the three-dimensional structural predictions fall within the unit cell accounts for the observed diffraction pattern.

[0059] In accordance with still another aspect of the present invention, there are provided methods for improving the accuracy of possible predicted possible protein structure(s), said methods comprising determining the degree to which predicted structures appropriately have experimentally determined fast amides on the surface thereof, and selecting predicted structures which most closely match the expected number and/or identity of fast amides on the surface thereof as more accurate models of protein structure. In a presently preferred embodiment, the identity of surface-located fast amides in a protein are experimentally determined by hydrogen exchange analysis.

[0060] In accordance with yet another aspect of the present invention, there are provided methods for selecting more accurate predicted protein structure(s) from among a plurality of predicted protein structure(s), said methods comprising determining the degree to which predicted structures appropriately have experimentally determined fast amides on the surface thereof, and selecting predicted structures which most closely match the expected number and/or identity of fast amides on the surface thereof as accurate models of protein structure.

[0061] Accordingly, the present invention provides methods for high-throughput protein structure determination and methods of selecting which of a plurality of calculated or predicted structures are most accurate based on comparisons with experimental hydrogen exchange profiles.

BRIEF DESCRIPTION OF THE FIGURES

[0062] **Figure 1** illustrates structure predictions for a CASP4 target protein ranked by the RMSD of the residuals between the COREX calculated rate fingerprint, and the COREX calculated structure rate fingerprint.

[0063] **Figure 2** illustrates structure predictions for a CASP4 target protein ranked in CASP4 by the number of correctly aligned residues with the crystal structure.

[0064] **Figure 3** illustrates structure predictions for a CASP4 target protein ranked by the RMSD of the residuals between the COREX calculated rate fingerprint, and the COREX calculated structure rate fingerprint.

[0065] **Figure 4** illustrates a summary of the 10-second deuteration results are shown for 21 Thermotoga proteins that were analyzed, whose amino acid lengths varied from 76 to 461 residues. Dark regions indicate fast exchanging amides ("fast amides") and clear regions indicate stretches of no exchange. Regions of four or more fast exchanging amides are circled.

[0066] **Figure 5** collectively illustrates TM0449 structure determination. **Figure 5A** depicts a ten-second amide hydrogen/ deuterium exchange map for TM0449. The horizontal bars are the protein's pepsin-generated fragments that had been produced, identified, and used as exchange rate probes in the subsequent 10-second deuteration study. The number of deuterons that went on to each peptide in 10 seconds is indicated by the number of grey residues in each peptide. Deuterium labeling was manually assigned to residue positions within the protein by first optimizing consensus in deuterium content of overlapping peptide probes, followed by further clustering of labeled amides together in the center of unresolved regions (with vertical bars indicating the range of possible location assignments), generating the consensus map at the top, in which two extensive segments are seen to be deuterium labeled: 1 (Phe 31-Glu 38) and 2 (Ser88-Lys 93). **Figure 5B** shows the electron density of the crystal indicates two regions of disordered sequence, corresponding to the segments 1 and 2. **Figures 5C** and **5D** show detailed electron density maps are shown , in which density is not visualized between the Phe 31 to Glu 39 and Ser 88 to Ser 95 regions of the TM0449 3-D structure. DXMS-determined disorder constitutes 6.4% of this protein's sequence.

[0067] **Figure 6** illustrates the on-exchange map of TM0505 and indicates three internal segments (A, B, and C) of rapidly exchanging amides. The internal segments are mapped onto the crystal structure of the GroES protein homolog of TM0505. The *M. tuberculosis* GroEL subunit is shown in dark grey and the heptamer complex of *M. tuberculosis* GroES subunits is shown in light gray. The homologous location of rapid exchange sites in the *T. maritima* protein are indicated in light grey. Disorder constitutes 16.3% of this protein's sequence.

[0068] **Figure 7** collectively shows a comparison of rate maps. **Figure 7A** shows TM1171 and **Figure 7B** shows TM0160, both showing substantial C-terminal disorder (circled sequences). Four truncated constructs of each protein were made by eliminating the C- terminal regions (D1-D4). C. **Figure 7C** shows that repeat DXMS analysis demonstrates that deletion constructs of TM0160 preserve the core full-length structure. Full-length TM0160, and its longest truncation (D3), were on-exchanged variously for 10, 100, 1,000, and 10,000 seconds at 0° C, exchange-quenched and subjected to comparative DXMS analysis as described herein. The resulting comprehensive exchange maps for full-length (Figure 7B) and D3 truncated (Figure 7C) had virtually identical patterns (10 second exchange time shown).

[0069] **Figure 8** collectively illustrates the exchange maps of the *Thermotoga maritima* proteins studies herein. Percentages indicate the amount of rapid exchange in amino acid segments of four or more residues, as a percentage of the entire sequence. **Figures 8A** and **8B** are proteins that crystallized and diffracted well. **Figures 8C-8E** are proteins that did not crystallize or had poor diffraction properties. Dark regions indicated fast exchanging amides and clear regions indicate stretches of no exchange. Regions of four or more fast exchanging amides are circled.

[0070] **Figure 9** illustrates a spectrin construct R1617 peptide map resulting from combined pepsin plus fungal protease XIII.

[0071] **Figure 10** shows the assignment of exchanging amides in spectrin R1617 into slow, medium, and fast-exchanging classes.

[0072] **Figure 11** illustrates the construction of low and high resolution exchange rate maps for spectrin construct R1617.

[0073] **Figure 12** illustrates a comparison of high resolution exchange rate maps obtained from DXMS data versus COREX analysis for spectrin construct R1617.

[0074] **Figure 13** shows examples of definition of Atomic units (AU) and setup for linear programming in the HR-DXMS deconvolution algorithm.

[0075] **Figure 14** illustrates the results of validation studies and the ability of HR-DXMS deconvolution algorithm and software to correctly calculate exchange rate profiles for simulated data derived from COREX analysis of a spectrin construct with and without introduced error.

[0076] **Figure 15** illustrates the results of validation studies and the ability of HR-DXMS deconvolution algorithm and software to correctly calculate exchange rate profiles for simulated data derived from NMR measurements of horse cytochrome c.

DETAILED DESCRIPTION OF THE INVENTION

[0077] The present invention constitutes a novel approach to structure determination that combines computational three-dimensional prediction methods with high- quality prediction-constraining information provided by experimentally acquired amide hydrogen exchange rate data for a protein, preferably acquired by amide hydrogen- deuterium exchange mass spectroscopy. This new approach will significantly accelerate the pace of protein structure elucidation.

[0078] Considerable theoretical and experimental evidence has established that the exchange rates of the peptide amide bond hydrogens within proteins are exquisitely dependent upon protein structure and thermodynamic stability. A number of enhancements to amide hydrogen/ deuterium exchange-mass spectrometry have recently been developed that allow the exchange rates of all of a protein's peptide amide hydrogens (its exchange rate "fingerprint") to be determined in days. Furthermore, the present invention presents the

development of computational approaches (the COREX algorithm) in combination with hydrogen exchange information that is capable of reliably predicting amide hydrogen exchange rate fingerprints from actual or presumed protein structure(s).

[0079] Three-dimensional determination of protein structure is required for a fundamental understanding of how proteins or protein modifications participate in human disease, and can provide a dramatically effective guide to the rational design of therapeutics to clinically important targets. The pressing need for this information in a timely manner contrasts with the agonizingly slow pace of present high-resolution structure determination methods. The present invention provides a solution with a new approach to protein structure that combines purely computational predictive methods with experimentally determined constraints of exceptional utility: peptide amide hydrogen/ deuterium exchange rate experimental data acquired by advanced mass spectrometric techniques (DXMS). A simple approach has been devised and termed the "DXMS-calculated rate - protein structure prediction validity filter" that allows such constraints, rapidly acquired by DXMS, to be directly applied to the refinement of the output of virtually any protein structure predictive method.

[0080] According to methods of the present invention, a protein or polypeptide's three dimensional structure is determined by performing hydrogen exchange measurements on the protein of interest to determine the amide hydrogen exchange rates for the majority of the amides in the protein, which together constitute its exchange rate "fingerprint". Optionally, the subset of these amides that are exchanging at the fastest possible rate are identified from this data, as they must be protein surface amino acid residues if they are to exchange at this maximal rate in a structured protein. Multiple possible three-dimensional (3-D) structures are proposed for the protein, employing any means available, including computational approaches using homology modeling, threading, and *ab initio* methods. In a preferred embodiment, the above noted hydrogen-exchange-derived identification of the identity of protein surface amino acid residues is used to refine the set of structural predictions made. The COREX algorithm, or other methods by which hydrogen exchange rates can be estimated from actual or proposed protein 3-D structures are used to calculate the virtual hydrogen exchange rate fingerprint for each of the several proposed structure(s) for the target

protein. These calculated fingerprints are compared to the true, experimentally determined rate fingerprint by employing methods such as root mean square deviation, or more advanced methods for such comparisons, and the structural predictions with calculated exchange rate fingerprints most closely matching experimentally determined fingerprints identified as the most accurate, or correct structural prediction.

[0081] As used herein, the phrase "protein structure prediction" refers to any method of estimating or approximating or determining the three-dimensional structure or model of a protein of interest. The methods of the present invention provide a novel method of assessing the degree to which such predictions match certain informative and readily accessible experimental measurements of the protein through the use of hydrogen exchange analysis. Hydrogen exchange analysis can be integrated into any known or novel methods of structure prediction available in the art. The present invention further provides methods wherein a hydrogen exchange rate map or fingerprint map of protein can be experimentally determined.

[0082] As used herein, the phrase "hydrogen exchange analysis" refers to any method by which measurement of the exchange rates of a peptide hydrogen with an isotope of hydrogen (for example, deuterium or tritium), present in the environment surrounding the protein (whether in soluble or crystalline form), are used to gain insight to the structure or stability of a protein as a whole, or portions or regions thereof. This includes both amide hydrogen exchange and alkyl hydrogen exchange. For more than 40 years, peptide amide hydrogen-exchange techniques have been employed to study the thermodynamics of protein conformational change and to probe the mechanisms of protein folding (see, *e.g.*, Englander and Englander, *Meth. Enzymol.* **232**:26-42, 1994; and Bai *et al., Meth. Enzymol.* **259**:344, 1995). More recently, they have proven to be increasingly powerful methods by which protein dynamics, domain structure, regional stability and function can be studied (see, *e.g.*, Englander *et al., Prot. Sci.* **6**:1101-1109, 1997). The principle of hydrogen-exchange reflects the fact that many hydrogens (commonly known as acidic hydrogens such as -OH, $-NH_2$, -SH, and peptide amide hydrogens) are not permanently attached to the protein, but continuously and reversibly interchange with hydrogen present in their external immediate environment. Most acidic hydrogen exchanges occur too rapidly to be experimentally useful. An important exception is the more slowly exchanging peptide amide hydrogen (main-chain

amide hydrogen) present in every amino acid except proline, thereby providing a way of examining protein structure and stability.

[0083] As used herein, the phrase "alkyl hydrogen exchange" refers to methods by which certain hydrogens on the side chains of a proteins amino acids can be induced to undergo exchange with heavy hydrogen in solvent water, as described by Anderson and Goshe.

[0084] The hydrogen exchange reaction can be experimentally followed by using tritiated or deuterated solvent. The chemical mechanisms of the exchange reactions are understood, and several well-defined factors can profoundly alter exchange rates. One of these factors is the extent to which a particular exchangeable hydrogen is exposed or accessible to solvent. The exchange reaction proceeds efficiently only when a particular peptide amide hydrogen is fully exposed to solvent. In a completely unstructured polypeptide chain, all peptide amide hydrogens are maximally accessible to water and exchange at their maximal possible rate, which is approximately (within a factor of 30) the same for all amides; a half-life of exchange in the range of one second at 0 °C and pH 7.0. Exact exchange rates expected for particular amide hydrogens in fully unstructured segments can be reliably calculated from knowledge of the temperature, pH and the primary amino acid sequence involved (see, *e.g.*, Molday *et al.*, *Biochemistry* **11**:150, 1972; and Bai *et al.*, *Proteins: Str. Funct. Gen.* **17**:74-86, 1993).

[0085] As used herein, "naturally occurring amino acid" and "naturally occurring R-group" includes L-isomers of the twenty amino acids naturally occurring in proteins. Naturally occurring amino acids are glycine, alanine, valine, leucine, isoleucine, serine, methionine, threonine, phenylalanine, tyrosine, tryptophan, cysteine, proline, histidine, aspartic acid, asparagine, glutamic acid, glutamine, arginine, and lysine. Unless specially indicated, all amino acids referred to in this application are in the L-form.

[0086] "Unnatural amino acid" and "unnatural R-group" includes amino acids that are not naturally found in proteins. Examples of unnatural amino acids included herein are racemic mixtures of selenocysteine and selenomethionine. In addition, unnatural amino acids include the D or L forms of, for example, nor-leucine, para-nitrophenylalanine, homophenylalanine, para-fluorophenylalanine, 3-amino-2-benzylpropionic acid, homoarginines, D-phenylalanine, and the like.

[0087] "R-group" refers to the substituent attached to the α-carbon of an amino acid residue. An R-group is an important determinant of the overall chemical character of an amino acid. There are nineteen natural R-groups found in proteins, which make up the twenty naturally occurring amino acids.

[0088] One of the twenty naturally occurring amino acids, glycine, is alpha unsubstituted and achiral. "α-carbon" refers to the chiral carbon atom found in an amino acid residue. Typically, four different substituents will be covalently bound to said α-carbon including an amine group, a carboxylic acid group, a hydrogen atom, and an R-group.

[0089] "Positively charged amino acid" and "positively charged R-group" includes any naturally occurring or unnatural amino acid having a positively charged side chain under normal physiological conditions. Examples of positively charged, naturally occurring amino acids include arginine, lysine, histidine, and the like.

[0090] "Negatively charged amino acid" and "negatively charged R-group" includes any naturally occurring or unnatural amino acid having a negatively charged side chain under normal physiological conditions. Examples of negatively charged, naturally occurring amino acids include aspartic acid, glutamic acid, and the like.

[0091] "Hydrophobic amino acid" and "hydrophobic R-group" includes any naturally occurring or unnatural amino acid having an uncharged, nonpolar side chain that is relatively insoluble in water. Examples of naturally occurring hydrophobic amino acids are alanine, leucine, isoleucine, valine, proline, phenylalanine, tryptophan, methionine, and the like.

[0092] "Hydrophilic amino acid" and "hydrophilic R-group" includes any naturally occurring or unnatural amino acid having a charged polar side chain that is relatively soluble in water. Examples of naturally occurring hydrophilic amino acids include serine, threonine, tyrosine, asparagine, glutamine, cysteine, and the like.

[0093] Modified forms of a protein of interest include forms having one or more R-group modifications to the amino acids of the parent protein or having a substitution of one or more amino acids, either conservative or non-conservative substitutions, that result in a modification of the protein amino acid sequence. For example, a modified form of a protein

will have an R-group on one or more α-carbon other than the prescribed arrangements of R-groups associated with one or more α-carbon of the parent protein. As used herein. A "conservative substitution" is an amino acid change that does not affect the three dimensional structure of the protein, as is known in the art, for example, substitution of a polar for a polar residue, a non-polar for a non-polar residue, *etc.*

[0094] Modifications and substitutions are not limited to replacement of amino acids. As used herein, "mutant", "mutated", "modified" or "daughter" forms of the protein of interest also include for example, deletion(s), replacement(s) or addition(s) of portions of the parent protein. For a variety of purposes, such as increased stability, solubility, or configuration concerns, one skilled in the art will recognize the need to introduce these and other such modifications. Examples of such other modifications include incorporation of rare amino acids, dextra-amino acids, glycosylation sites, cytosine for specific disulfide bridge formation, and the like. The modified peptides can be chemically synthesized, or the isolated gene can be subjected to site-directed mutagenesis, or a synthetic gene can be synthesized and expressed in bacteria, yeast, baculovirus, tissue culture, and so on.

[0095] Modified forms of the proteins contemplated for use in the practice of the present invention may be prepared in a number of ways available to the skilled artisan. For example, the gene encoding a parent protein may be mutated or modified at those sites identified by the hydrogen exchange methods described herein as corresponding to amino acid residues in unstructured areas by means currently available to the artisan skilled in molecular biological techniques. Such techniques include oligonucleotide-directed mutagenesis, deletion, chemical mutagenesis, and the like. The protein encoded by the mutant gene is then produced by expressing the gene in, for example, a bacterial, mammalian, insect or plant expression system.

[0096] Alternatively, modified forms may be generated by site specific-replacement of a particular amino acid with an unnaturally occurring amino acid or mimetic. As such, modified forms may be generated through replacement of an amino acid residue or a particular cysteine or methionine residue with selenocysteine or selenomethionine. This may be achieved by growing a host organism capable of expressing either the wild-type or mutant polypeptide on a growth medium depleted of natural cysteine or methionine or both and

growing on medium enriched with either selenocysteine, selenomethionine, or both. These and similar techniques are described in Sambrook *et al.*, (Molecular Cloning, A Laboratory Manual, 2nd Ed. (1989) Cold Spring Harbor Laboratory Press).

[0097] Another suitable method of creating modified forms of a protein for use in the methods of the present invention is based on a procedure described in Noel and Tsai, *J. Cell. Biochem.*, **40**:309-320, 1989. In so doing, the nucleic acids encoding the protein can be synthetically produced using oligonucleotides having overlapping regions, said oligonucleotides being degenerate at specific bases so that mutations are induced.

[0098] In designing the nucleic acid sequences to encode a protein of interest, it may be desirable to reengineer the gene for improved expression in a particular expression system. For example, it has been shown that many bacterially derived genes do not express well in plant systems. In some cases, plant-derived genes do not express well in bacteria. This phenomenon may be due to the non-optimal G+C content and/or A+T content of said gene relative to the expression system being used. For example, the very low G+C content of many bacterial genes results in the generation of sequences mimicking or duplicating plant gene control sequences that are highly A+T rich. The presence of A+T rich sequences within the genes introduced into plants (*e.g.*, TATA box regions normally found in promoters) may result in aberrant transcription of the gene(s). In addition, the presence of other regulatory sequences residing in the transcribed mRNA (*e.g.* polyadenylation signal sequences (AAUAAA) or sequences complementary to small nuclear RNAs involved in pre-mRNA splicing) may lead to RNA instability. Therefore, one goal in the design of genes is to generate nucleic acid sequences that have a G+C content that affords mRNA stability and translation accuracy for a particular expression system.

[0099] Due to the plasticity afforded by the redundancy of the genetic code (*i.e.*, some amino acids are specified by more than one codon), evolution of the genomes of different organisms or classes of organisms has resulted in differential usage of redundant codons. This "codon bias" is reflected in the mean base composition of protein coding regions. For example, organisms with relatively low G+C contents utilize codons having A or T in the third position of redundant codons, whereas those having higher G+C contents utilize codons having G or C in the third position. Therefore, in reengineering genes for expression, one

may wish to determine the codon bias of the organism in which the gene is to be expressed. Looking at the usage of the codons as determined for genes of a particular organism deposited in GenBank can provide this information. After determining the bias thereof, the new gene sequence can be analyzed for restriction enzyme sites as well as other sites that could affect transcription such as exon:intron junctions, polyA addition signals, or RNA polymerase termination signals.

[0100] Genes encoding the protein of interest can be placed in an appropriate vector and can be expressed using a suitable expression system. An expression vector, as is well known in the art, typically includes elements that permit replication of said vector within the host cell and may contain one or more phenotypic markers for selection of cells containing the gene. The expression vector will typically contain sequences that control expression such as promoter sequences, ribosome binding sites, and translational initiation and termination sequences. Expression vectors may also contain elements such as subgenomic promoters, a repressor gene or various activator genes. The artisan may also choose to include nucleic acid sequences that result in secretion of the gene product, movement of said product to a particular organelle such as a plant plastid (see U.S. Patent Nos. 4,762,785; 5,451,513 and 5,545,817, which are each incorporated herein by reference in their entirety) or other sequences that increase the ease of peptide purification, such as an affinity tag.

[0101] A wide variety of expression control sequences are useful in expressing native/parent or modified forms of the protein of interest when operably linked thereto. Such expression control sequences include, for example, the early and late promoters of SV40 for animal cells, the *lac* system, the *trp* system, major operator and promoter systems of phage S, and the control regions of coat proteins, particularly those from RNA viruses in plants. In *E. coli*, a useful transcriptional control sequence is the T7 RNA polymerase binding promoter, which can be incorporated into a pET vector as described by Studier *et al.*, *Methods Enzymology* **185**:60-89, 1990.

[0102] For expression, a desired gene should be operably linked to the expression control sequence and maintain the appropriate reading frame to permit production of the desired protein or modified form thereof. Any of a wide variety of well-known expression vectors are of use in the methods of the present invention. These include, for example, vectors

comprising segments of chromosomal, non-chromosomal and synthetic DNA sequences such as those derived from SV40, bacterial plasmids including those from *E. coli* such as col E1, pCR1, pBR322 and derivatives thereof, pMB9, wider host range plasmids such as RP4, phage DNA such as phage S, NM989, M13, and other such systems as described by Sambrook *et al.*, (Molecular Cloning, A Laboratory Manual, 2nd Ed. (1989) Cold Spring Harbor Laboratory Press), which is incorporated by reference herein.

[0103] A wide variety of host cells are available for expressing mutants of the present invention. Such host cells include, for example, bacteria such as *E. coli*, *Bacillus* and *Streptomyces*, fungi, yeast, animal cells, plant cells, insect cells, and the like.

[0104] "Purified" or "isolated" refers to a protein or nucleic acid that has been separated from its natural environment. Contaminant components of its natural environment may include enzymes, hormones, and other proteinaceous or non-proteinaceous solutes. In one embodiment, the isolated molecule, in the case of a protein, will be purified to a degree sufficient to obtain at least 15 residues of N-terminal or internal amino acid sequence or to homogeneity by SDS-PAGE under reducing or non-reducing conditions using Coomassie blue or silver stain. In the case of a nucleic acid the isolated molecule will preferably be purified to a degree sufficient to obtain a nucleic acid sequence using standard sequencing methods.

[0105] By a "substantially pure polypeptide" or "substantially pure protein" is meant a polypeptide or protein which has been separated from components which naturally accompany it.

[0106] Typically, the polypeptide is substantially pure when it is at least 60%, by weight, free from the proteins and naturally-occurring organic molecules with which it is naturally associated. Preferably, the preparation is at least 75%, more preferably at least 90%, and most preferably at least 99%, by weight, polypeptide. A substantially pure protein or polypeptide may be obtained, for example, by extraction from a natural source; by expression of a recombinant nucleic acid encoding a polypeptide; or by chemically synthesizing the protein. Purity can be measured by any appropriate method (*e.g.*, column chromatography, polyacrylamide gel electrophoresis, or by HPLC analysis).

[0107] "Degenerate variations thereof" refers to changing a gene sequence using the degenerate nature of the genetic code to encode proteins having the same amino acid sequence yet having a different gene sequence. Degenerate gene variations thereof can be made encoding the same protein due to the plasticity of the genetic code, as described herein.

[0108] "Expression" refers to transcription of a gene or nucleic acid sequence, stable accumulation of nucleic acid, and the translation of that nucleic acid to a polypeptide sequence. Expression of genes also involves transcription of the gene to make RNA, processing of RNA into mRNA in eukaryotic systems, and translation of mRNA into proteins. It is not necessary for the genes to integrate into the genome of a cell in order to achieve expression. This definition in no way limits expression to a particular system or to being confined to cells or a particular cell type and is meant to include cellular, transient, *in vitro, in vivo*, and viral expression systems in both prokaryotic, eukaryotic cells, and the like.

[0109] "Foreign" or "heterologous" genes refers to a gene encoding a protein whose exact amino acid sequence is not normally found in the host cell.

[0110] "Promoter" and "promoter regulatory element", and the like, refers to a nucleotide sequence element within a nucleic acid fragment or gene that controls the expression of that gene. These can also include expression control sequences. Promoter regulatory elements, and the like, from a variety of sources can be used efficiently to promote gene expression. Promoter regulatory elements are meant to include constitutive, tissue-specific, developmental-specific, inducible, subgenomic promoters, and the like. Promoter regulatory elements may also include certain enhancer elements or silencing elements that improve or regulate transcriptional efficiency. Promoter regulatory elements are recognized by RNA polymerases, promote the binding thereof, and facilitate RNA transcription.

[0111] "Structure coordinates" refers to Cartesian coordinates (x, y, and z positions) derived from mathematical equations involving Fourier synthesis as determined from patterns obtained via diffraction of a monochromatic beam of X-rays by the atoms (scattering centers) of a polypeptide in crystal form. Diffraction data are used to calculate electron density maps of repeating protein units in the crystal (unit cell). Electron density maps are used to establish the positions of individual atoms within a crystal's unit cell. The term "crystal

structure coordinates" refers to mathematical coordinates derived from mathematical equations related to the patterns obtained on diffraction of a monochromatic beam of X-rays by the atoms (scattering centers) of a polypeptide in crystal form. The diffraction data are used to calculate an electron density map of the repeating unit of the crystal. The electron density maps are used to establish the positions of the individual atoms within the unit cell of the crystal.

[0112] The term "selenomethionine substitution" refers to the method of producing a chemically modified form of the crystal of a protein. The protein is expressed by bacteria in media that is depleted in methionine and supplemented with selenomethionine. Selenium is thereby incorporated into the crystal in place of methionine sulfurs. The location(s) of selenium are determined by X-ray diffraction analysis of the crystal. This information is used to generate the phase information used to construct a three-dimensional structure of the protein.

[0113] "Heavy atom derivatization" refers to a method of producing a chemically modified form of a crystal. In practice, a crystal is soaked in a solution containing heavy atom salts or organometallic compounds, *e.g.*, lead chloride, gold thiomalate, thimerosal, uranyl acetate, and the like, which can diffuse through the crystal and bind to the protein's surface. Locations of the bound heavy atoms can be determined by X-ray diffraction analysis of the soaked crystal. This information is then used to construct phase information which can then be used to construct three-dimensional structures of the enzyme as described in Blundel, T. L., and Johnson, N. L., Protein Crystallography, Academic Press (1976), which is incorporated herein by reference.

[0114] "Unit cell" refers to a basic parallelepiped shaped block. Regular assembly of such blocks may construct the entire volume of a crystal. Each unit cell comprises a complete representation of the unit pattern, the repetition of which builds up the crystal. "Space group" refers to the arrangement of symmetry elements within a crystal.

[0115] "Molecular replacement" refers to generating a preliminary model of a protein whose structural coordinates are unknown, by orienting and positioning a molecule whose structural coordinates are known within the unit cell of the unknown crystal so as best to

account for the observed diffraction pattern of the unknown crystal. Phases can then be calculated from this model and combined with the observed amplitudes to give an approximate Fourier synthesis of the structure whose coordinates are unknown. This in turn can be subject to any of the several forms of refinement to provide a final, accurate structure of the unknown crystal (Lattman, E., *Meth. Enzymol.* 11:55-77, 1985; Rossmann, MG., ed., "The Molecular Replacement Method" 1972, Int, Sci. Rev. Ser., No. 13, Gordon & Breach, New York).

[0116] The term "protein" or "polypeptide" is used herein in a broad sense which includes, for example, polypeptides and oligopeptides, and derivatives thereof, such as glycoproteins, lipoproteins, and phosphoproteins, and metalloproteins. The essential requirement is that the protein contains one or more peptide (--NHCO--) bonds, as the amide hydrogen of the peptide bond (as well as in the side chains of certain amino acids) has certain properties which lends itself to analysis by proton exchange. The protein may be identical to a naturally occurring protein, or it may be a binding fragment or mutant of such a protein. The fragment or mutant may have the same or different binding characteristics relative to the parent protein.

[0117] The numerous small peptide fragments that are produced and analyzed by the methods of the present invention are likely to all be in random coil configuration: they are small, with little opportunity for structure-forming interactions, and are continuously contacted with several structure-breaking denaturants. According to certain invention methods, deuterated proteins are shifted to slowed exchange conditions (that include a very acidic pH), admixed with denaturing guanidinium salts, optionally disulfide-reduced, subject to proteolysis to generate a population of small fragments, and then admixed with acetonitrile, again under very acid conditions. As these fragments are in random coil configuration, the rates of exchange of each amide, in each peptide, under the slowed exchange ("quench") conditions as employed herein can be calculated from a knowledge of the amino acid sequence of each fragment (Bai *et al., supra*) as well as determined experimentally by fragmentation-LC-MS analysis of initially equilibrium-deuterated protein or peptides. As demonstrated herein, such calculations and measurements are employed to provide precise corrections for deuterium losses from peptides that occur in the course of the

analysis, and to provide an adjunctive method for further localizing deuterium on peptide amides, when the fragmentation data alone is insufficient to achieve the desired resolution.

[0118] The protein of interest is first labeled under conditions wherein native hydrogens are replaced by the isotope of hydrogen (this is the "on-exchange" step). The reaction conditions are then altered to slowed hydrogen exchange conditions, or exchange "quench" conditions for further analysis of exchange rates. The phrase "slowed hydrogen exchange conditions" as used herein, refers to conditions where the rate of exchange of normal hydrogen for an isotope of hydrogen at amide hydrogens freely exposed to solvent is reduced substantially, *i.e.*, enough to allow sufficient time to determine, by the methods described herein, exchange rates and the location of amide hydrogen positions which had been labeled with heavy hydrogen. The hydrogen exchange rate is a function of such variables as temperature, pH and solvent, in addition to protein structure. The rate is decreased three fold for each 10 °C drop in temperature. In water, the minimum hydrogen exchange rate is at a pH of 2-3. The use of a temperatures in the range of about 0 - 10 °C, and a pH in the range of about 2-3 is preferred. Most presently preferred are conditions of about 0 °C and pH 2.2. As conditions diverge from the optimum pH, the hydrogen exchange rate increases, typically by 10-fold per pH unit increase or decrease away from the minimum. Use of high concentrations of a polar, organic cosolvent shifts the pH min to higher pH, potentially as high as pH 6 and perhaps, with certain solvents, even higher.

[0119] At pH 2.2 and 0 °C, the typical half life of a deuterium label at an amide position freely exposed to solvent water is about 70 minutes. Preferably, the slowed conditions of the present invention result in a half-life of at least 10 minutes, more preferably at least 60 minutes.

[0120] To achieve labeling of the protein of interest, the protein is incubated in buffer supplemented with deuterated water (preferably ^2H_2O), preferably of high concentration, preferably greater than 25% mole fraction deuterated water. This results in the time dependent reversible incorporation of deuterium label into every peptide amide on the surface of the protein through the mechanism of hydrogen exchange. These amides are referred to herein as "solvent accessible". A suitable buffer is phosphate buffered saline (PBS; 0.15 mM NaCl, 10 mM PO_4 (pH 7.4)). The use of small incubation volumes (about 0.1 - 10 μl)

containing high concentrations of protein (about 2 - 10 mg/ml) is preferred. This can be done, for example, by adding protein and buffer together in a tube, or by injecting an aliquot of protein solution into a flowing stream of isotope-containing buffer in a manner that results in the rapid mixing of the converging streams.

[0121] It is not necessary that the hydrogen exchange analysis rely on only a single choice of "on-exchange" time. Rather, the skilled worker may carry out the experiment using a range of on-exchange times, preferably spanning several orders of magnitude (seconds to days) to allow selection of on-exchange times which allow efficient labeling of the various peptide amides present in the protein, and at the same time minimize background labeling of other amide positions after off-exchange is completed.

[0122] In general, comparisons of the exchange behavior of alternative forms of a protein can be performed by either : (i) on-exchanging, in parallel, each of the forms of the protein, quenching exchange, performing localization studies on each form of the protein, and then comparing the deuteration patterns seen across the set of protein forms; and (ii) on-exchanging one form of the protein, transforming the protein to its alternative form (for example, inducing a conformation change, binding a ligand, *etc.*) and then off-exchanging the protein, said off-exchange terminated by quenching exchange. In both methods of analysis, the ratio of the exchange rates observed at any amide position is termed its exchange "protection factor", and this ratio is related to the change in free energy ("delta G") in the atomic environs of said amide by the relationship Delta G = -T ln (protection factor).

[0123] For off-exchange, the labeled protein is transferred to physiologic buffers identical to those employed during on-exchange, but which are substantially free of isotope. The incorporated isotopic label on the protein then exchanges off the protein at rates identical to its on-exchange rate everywhere except at amides which have been slowed in their exchange rate, for example, by virtue of the interaction of protein with a binding partner, or by conformational change.

[0124] In general, off-exchange is allowed to proceed for 2 to 20 times, more preferably about 10 times longer than the on-exchange period, as this allows off-exchange from the protein of greater than 99% of the on-exchanged isotope label.

[0125] In preferred embodiments, the off-exchange procedure may be accomplished by use of perfusive HPLC supports that allow rapid separation of peptide/protein from solvent (*e.g.*, Poros™ columns, PerSeptive Biosystems, Boston, Mass.), or by simple dilution into undeuterated solvent.

[0126] Determination of amide exchange rates in proteins requires performing studies across a broad range of on and off- exchange intervals. For brief on- and off-exchange intervals (1-2 minutes or less), the time necessary for binding protein to be applied to the matrix- containing column and both bind to binding partner and start off- exchange may be excessively long with the above approach. While the above approach will work well with on and off- exchange intervals as short as 1-2 minutes, limits to the ability of support matrices to promote the rapid molecular interaction of binding protein with binding partner will make study of exchange intervals shorter than this problematic with the above approach. While homogenous liquid phase reactions between a receptor and ligand may be quite fast (less than $1/10^{th}$ of a second), if one of the pair has been previously attached to a surface, then limitations to "transport processes" can substantially slow the binding interaction (to seconds).

[0127] To overcome this difficulty, the following modified approach is utilized for study of brief exchange intervals. Binding protein is contacted with isotope-containing solvent as above, but at the end of the desired on-exchange interval, the solution is contacted with a small volume of liquid phase binding partner. As both binding components are in homogenous liquid phase, complex formation occurs at intervals well less than one second. An excess of aqueous solvent devoid of heavy hydrogen is then optionally added to the binding protein- binding partner complex mixture to effect a substantial dilution (1/10 to 1/1000, preferably 1/100) of the isotope in the mixture, thereby initiating off-exchange. This mixture is then rapidly applied to a support matrix column (preferably by the flowing stream method) that is capable of binding and attaching the binding partner by any of a variety of methods that are operative at physiologic pH, including the avidin-biotin interaction (in this case the binding partner having been previously biotinylated and the matrix support bearing previously attached avidin) or by way of other well-characterized binding pair interactions.

[0128] Continued flow of solvent without isotope over the binding protein-binding partner-bound support matrix further initiates off-exchange. At the end of off-exchange, binding protein is then eluted and removed from the column with an appropriate buffer capable of dissociating the binding protein-binding partner complex; the binding partner-solid support interaction; or both. Preferably one employs procedures that are capable of selectively disrupting the binding protein-binding partner complex without disrupting the support matrix-binding partner interaction (for example, the avidin-biotin interaction) as this will result in the preferred specific elution and recovery from the column of pure off-exchanged binding protein, unadulterated with confounding binding partner.

[0129] A preferred embodiment employs binding protein that is first contacted with isotope-containing solvent, and, at the end of the desired on-exchange interval, this solution is contacted with a solution of a previously biotinylated binding partner, with such prior biotinylation being accomplished by any of a number of well known procedures. Complex formation between biotinylated binding partner and binding protein is allowed to occur, generally being complete in less than a second, and then this mixture is optionally diluted to initiate off-exchange, and injected into a flowing stream of physiologic aqueous solvent flowing over a column of support matrix consisting of avidin covalently bound to the matrix. The avidin utilized may variously consist of streptavidin, egg white avidin, or monomeric avidin, or other modified forms of avidin. The linkage to matrix may be by way of any of a variety of functionalities including sodium cyanoborohydride-stabilized Schiff base or that resulting from the cyanogen bromide procedure as applied to carbohydrate matrices. The solid matrices may consist of cross-linked agarose particles or preferably perfusive supports such as those (Poros products) provided by the Perceptive Biosystems company (solid support 20-AL and the like).

[0130] For many binding pairs off-exchange may be terminated and selective elution of binding protein accomplished by simply shifting pH to about 2.2 at 0 °C. These conditions disrupt many types of binding protein-binding partner complexes but do not disrupt the avidin-biotin interaction, thereby allowing retention on the column of biotinylated binding partner. If shifting to acidic conditions by itself does not result in elution of a particular binding protein, then one of a variety of additional denaturants can be added to the elution

solvent, including urea, guanidine hydrochloride, and guanidine thiocyanate at concentrations (preferably 2 - 4 M guanidine hydrochloride, 1 - 2 M guanidine thiocyanate) sufficient to elute binding protein but not at the same time disrupt the avidin-biotin interaction and thereby co-elute the binding partner. In general, these conditions do not disrupt the avidin-biotin interaction, even at room temperature. Finally, as above, reductants, such as TCEP, can optionally be admixed with the elution solvent so that it will be present in the binding partner sample when desired.

[0131] An additional advantage of the support matrix approach to exchange reactions is that certain embodiments require that the binding protein and binding partner of interest be on-exchanged, complexed with each other, and off-exchanged while present within a mixture of other proteins and biomolecules. In these embodiments, as off-exchange proceeds, it is necessary to isolate the specific binding pair complex of interest. In a preferred embodiment this is accomplished with support matrices as follows. Previously biotinylated binding partner is contacted with a sample containing a mixture of proteins, perhaps a suspension of intact, living cells, or a whole cell extract or digest, or a biologic fluid, such as serum, plasma or blood that also contains the binding protein of interest. Said contacting and mixing results in formation of the biotinylated binding partner-binding protein complex. This mixture, of which the binding pair may be a minor component, is then passed over the aforementioned support matrix containing avidin, wherein the biotinylated complex of interest will specifically attach to the matrix. Washing of the support with aqueous solvent continues (or when desired may initiate) off-exchange and removes from the matrix the irrelevant proteins that were present in the initial mixture, and thereby purifies the binding protein- binding partner complex. At the end of the off-exchange interval, the purified binding protein is simultaneously eluted and shifted to slow exchange conditions as above with an aliquot of appropriate eluent.

[0132] Certain target proteins require lipid or detergent environments for expression of their physiologic structure and function. Slowed-exchange-compatible proteolysis of such protein targets can be accomplished with current methods, but further analysis (c18 reversed-phase chromatography, ESI-MS) is not possible because of interference from the associated lipids and/or detergents. The use of microfluidic devices allows such interfering substances

to be efficiently and rapidly separated from the peptide fragments, allowing their effective analysis, for example using deuterium exchange-mass spectrometry (DXMS).

[0133] Through the use of microfluidic devices, solutions containing target proteins have their buffer composition changed by allowing effective diffusion of the smaller buffer components (^2H_2O, H_2O, salts, ligands) without effective diffusion of the target protein. In one embodiment, small regenerated cellulose microdialysis fibers (13,000 or 18,000 MWCO , approximately 200u ID; Spectrum Inc.) are encased in PEEK tubing (15/1000 inch ID) with end fittings that allow a countercurrent sheath solvent flow of exchange solvent while the protein solution flows through the microdialysis fiber. Such devices are capable of very efficient ^2H_2O exchange in short times, for example, effecting change to 95% ^2H_2O in three seconds at room temperature. Typical flow rates to achieve this end consist of 50 µl/minute for protein solution and 1000 µl/minute for sheath solution.

[0134] Such microfluidic devices can also be used to semipurify peptide mixtures that are contaminated with interfering lipids and detergents, such as proteolytic digests of membrane protein preparations. In this application, the proteolytic digest of such a protein is passed through the bore of the microdialysis fiber (flow 5-50 µl/minute) while the countercurrent sheath flow (100 - 400 µl/min), into which peptide fragments can transfer, (but not the more slowly diffusing and non-dializable lipid/detergent micelles), is directed to and collected on the c18 column for subsequent acetonitrile- gradient elution and MS. The result is that the digest peptides can be analyzed without interference from the lipid/detergent.

[0135] Non-constrained devices which utilize differential diffusion to effect changes in buffer composition (such as the "H- reactor" patented by Micronics, Inc.) can also be employed for these purposes. With these devices, flow of sample and exchange buffer is concurrent, not countercurrent, and exchange is therefore necessarily less efficient for a given volume of exchange buffer employed.

Protein Fragmentation Methods

A. Improved proteolysis fragmentation

[0136] In one preferred method of hydrogen exchange analysis, improved proteolysis fragmentation is employed. In this improved proteolysis method, a simple endopeptidase proteolysis is used to generate a dense sequence-overlapping population of protein fragments for analysis. Prior teachings had found that the common acid-resistant endopeptidases alone, such as pepsin, were not useful in highly localizing amide hydrogen exchange due to insufficient ability to fragment target proteins under acceptable slowed exchange conditions. Pepsin, as employed in the prior art typically had generated a relatively small number of fragments, generally 10-25 amino acids long. The label incorporated on these few useable pepsin-generated peptides was then used to infer the location of label, at best localizing within a range of about 10-25 amino acids. Subsequent art taught the use of acid-resistant carboxypeptidases (progressive degradation) after an initial employment of endopeptidases, to localize the labeled amino acid positions within peptides generated when a detailed resolution, such as within 1-5 amino acid residues, is desired.

[0137] In accordance with the present invention, improved methods that dramatically speed proteolysis, and modulate the sites and patterns of proteolysis by endoproteinases are employed so as to produce highly varied and highly efficient fragmentation of the labeled protein in a single step, thereby avoiding the use of carboxypeptidases completely, an improvement which simplifies the fragmentation and affords a considerable savings of time and cost. While these improvements work best in combination with each other, they can be grouped into 3 categories: (i) use of denaturants (systematically varying the type, concentration, duration of denaturation, type of endoproteinase)s) employed, and the duration of endopeptidase digestion) to greatly speed proteolysis and modulate the resulting pattern of fragmentation ; (ii) use of solid-state proteolysis with acid- resistant endopeptidases selected for their efficiency and distinctive fragmentation preferences with respect to each other under optimal quench conditions; and (iii) use of water-soluble phosphines to effect rapid and efficient disulfide reduction under quench conditions

[0138] The use of such endopeptidases under optimized conditions described herein routinely results in the generation of a population of endopeptidase-generated fragments substantially spanning the full length of the majority of proteins studied to date, and, as importantly, yields a large number of additional peptides that partially and mutually overlap in sequence with each other, all obtainable in useful yield. Preferably, the population of fragments contains sequence-overlapping fragments wherein more than half, more preferably 60% - 80%, of the members of the population have sequences that are overlapped by the sequences of other members by all but 1-5 amino acid residues. In addition, it is preferable that a majority of members of the population of fragments is present in an analytically sufficient quantity to permit its further characterization, for example, by LC-MS analysis.

[0139] An example of the application of this improved proteolysis method and the power of deuterium exchange-mass spectrometry (DXMS) to elucidate protein structure and organization can be found in Hamuro *et al., J. Mol. Biol.* **321**:703-714, 2002. Additional references include Hamuro and Woods, *J. Cell. Biochem.*, **37**:89-98, 2001; Hamuro *et al., J. Mol. Biol.* **323**:871-881, 2002; Hamuro *et al., J. Mol. Biol.* **327**:1065-1076, 2003; Englander *et al., Proc. Natl. Acad. Sci. USA* **100**:7057-7062, 2003; and Zawadzki *et al., Protein Sci.* **12**:1980-1990, 2003.

B. Progressive proteolysis fragmentation

[0140] In another preferred method of hydrogen exchange analysis, progressive proteolysis (as defined above) is employed to produce protein fragments for label localization. The protein is subjected to a first fragmentation, *e.g.*, with an acid stable proteolytic enzyme, *e.g.*, an endopeptidase such as, for example, pepsin, under slow hydrogen exchange conditions to generate protein fragments. Following the first fragmentation, the resolution of the isotopic hydrogen labeled amides is equivalent to the protein fragment size. Finer localization of the labels is achieved by analysis of subfragments of the protein fragments, which subfragments are generated by progressive degradation of each isolated, labeled protein fragment under slowed exchange conditions. Alternatively, if the protein is smaller than about 30 kDa, the intact protein may be subjected to progressive degradation. For the purpose of the present invention, a protein or a protein fragment is said to be "progressively" (or "stepwise" or "sequentially") degraded if a series of fragments are

obtained which are similar to the series of fragments which would be achieved using an ideal exopeptidase, as defined and described in U.S. Patent No. 6,291,189, column 7, line 58 through column 8, line 33. An ideal exopeptidase will only remove a terminal amino acid. Thus, if the n amino acids of a protein fragment were labeled A_1 to A_n (the numbering starting at the terminus at which the degradation occurs), the series of subfragments produced by an ideal exopeptidase would be $A_2\sim\sim\sim A_n$, $A_3\sim\sim\sim A_{n-1}$—A_n, and finally A_n.

[0141] Preferably each subfragment of the series of subfragments obtained is shorter than the preceding subfragment in the series by a single terminal amino acid residue. However, it is to be understood that exopeptidases do not necessarily react in an ideal manner. Thus, for purposes of the present invention, a protein fragment is said to be progressively degraded, if the series of subfragments generated thereby is one wherein each subfragment in the series is composed of about 1-5 fewer terminal amino acid residues from one end than the preceding subfragment in the series, with preservation of the common other end of the subfamily members. The analyses of the successive subfragments are correlated in order to determine which amino acids of the parent protein fragment were isotopically labeled.

Protein fragmentation

[0142] When the progressive proteolysis protein fragmentation method is employed, the protein is subjected to acid proteolysis with high concentrations of at least one protease that is stable and proteolytically active in the aforementioned slowed hydrogen exchange conditions, *e.g.*, a pH of about 2 - 3, and a temperature of about 0 – 4 ° C, followed by C-terminal subfragmentation with an acid resistant carboxypeptidase, or N-terminal degradation with an acid resistant aminopeptidase. Suitable proteases for the first step include, for example, pepsin (Rogero *et al.*, *Meth. Enzymol.* **131**:508-517, 1986.), cathepsin-D (Takayuki *et al.*, *Meth. Enzymol.* **80**:565-581, 1981) Aspergillus proteases (Krishnan *et al.*, *J. Chromatography* **329**:165-170, 1985; Xiaoming *et al.*, *Carlsberg Res. Commun.* **54**:241-249, 1989; Zhu *et al.*, *App. Envir. Microbiol.* **56**:837-843, 1990), thermolysin (Fusek *et al.*, *J. Biol. Chem.* **265**:1496-1501, 1990) and mixtures of these proteases. In one preferred embodiment, pepsin is used, preferably at a concentration of 10 mg/mL pepsin at a temperature of about 0 °C and a pH of about 2.7 for about 5-30 minutes, preferably about 10 minutes.

Separation of protein fragments

[0143] In one embodiment of the invention, proteolytically fragmented, isotopic hydrogen-labeled protein fragments are separated prior to progressive degradation by means capable of resolving the protein fragments. Preferably, separation is accomplished by reverse phase high performance liquid chromatography (RP-HPLC) utilizing one or more of a number of potential chromatographic supports including C_4, C_{18}, phenol and ion exchange, preferably C_{18}.

[0144] Separating the isotopically labeled fragments from the many unlabeled peptides generated by fragmentation of the protein is done under conditions which minimize off-exchange of isotopic hydrogen from the labeled amide sites of the protein fragments. Small protein fragments have little secondary structure, thus amide hydrogens therein freely exchange with hydrogen from the solvent. Conditions for proteolysis and protein fragment separation must therefore be adjusted to slow off-exchange of isotopic hydrogen in order for the isotopic label to remain in place for a time sufficient to complete the method.

[0145] The RP-HPLC separation is preferably performed at a pH of about 2.1-3.5 and at a temperature of about 0 - 4.0 ° C, more preferably, at a pH of about 2.7 and at a temperature of about 0 ° C. The preferred separation conditions may be generated by employment of any buffer systems which operate within the above pH ranges, including, for example, citrate, phosphate, and acetate, preferably phosphate. Protein fragments are eluted from the reverse phase column using a gradient of similarly buffered polar co-solvents including methanol, dioxane, propanol, and acetonitrile, preferably acetonitrile. Eluted protein fragments are detected, preferably by ultraviolet light absorption spectroscopy performed at frequencies between about 200 and about 300 nM, preferably about 214 nM. The isotopic label is detected in a sampled fraction of the HPLC column effluent, preferably via either scintillation counting for a tritium label or by mass spectrometry for a deuterium label.

[0146] Acid proteases in general have broad cleavage specificity. Thus, they fragment the protein into a large number of different peptides. RP-HPLC resolution of co-migrating multiple peptides is substantially improved by employing a two-dimensional RP-HPLC separation. Preferably, the two sequential RP-HPLC separations are performed at

substantially different pH's, for example, a pH of about 2.7 for one separation and about 2.1 for the other sequential separation.

[0147] HPLC fractions from a first separation, containing isotopically labeled protein fragments, are then optionally subjected to a second dimension RP-HPLC separation. The second separation may be performed at a pH of from about 2.1 to about 3.5 and at a temperature of from about 0 to about 4° C, more preferably, at a pH of about 2.1 and at a temperature of about 0° C. The pH conditions for the chromatographic separation are maintained by employing a buffer system which operates at this pH, including citrate, chloride, acetate, phosphate, more preferably TFA (0.1-0.115%). Protein fragments are eluted from their reverse phase column with a similarly buffered gradient of polar co-solvents including methanol, dioxane, propanol, more preferably acetonitrile. Eluted protein fragments are detected, the content of isotopic label is measured, and labeled peptides identified as in the first HPLC dimension described above. Labeled protein fragments are isolated by collection of the appropriate fraction of column effluent. Elution solvents are removed by evaporation. The remaining purified protein fragments are each characterized as to primary amino acid structure by conventional techniques such as, for example, amino acid analysis of complete acid hydrolysates or gas-phase Edman degradation microsequencing. The location of the labeled protein fragments within the primary sequence of the intact protein may then be determined by referencing the previously known amino acid sequence of the intact protein. Residual phosphate frequently interferes with the chemical reactions required for amino acid analysis and Edman degradation. This interference is eliminated by the use of trifluoroacetic acid (TFA) in the second dimension buffer so that no residual salt, *i.e.*, phosphate remains after solvent evaporation.

[0148] In one embodiment, proteolytically fragmented, isotopic hydrogen-labeled protein fragments are first separated at pH 2.7 in phosphate buffered solvents and each eluted fragment peak fraction which contains isotopically-labeled amides is identified, collected, and then subjected to a second HPLC separation performed in TFA-buffered solvents at pH 2.1.

<u>High Resolution Sublocalization of Labeled Amides Within Label-Bearing Protein Fragments</u>

1. Subfragmentation of Protein Fragments

[0149] To localize an isotopic hydrogen labeled peptide amide to the single amino acid level, substantially every peptide bond within a purified label-bearing protein fragment is systematically cleaved. Acidic conditions must be used for this proteolysis because the small protein fragments and subfragments generated have no stable conformational structure and rapid loss of isotopic hydrogen label from the amides would occur if rates of exchange were not slowed by ambient acidic pH.

[0150] Progressive degradation is preferably achieved by treatment with at least one acid stable exopeptidase enzyme, more preferably with at least one carboxypeptidase. The progressive degradation is performed at acidic pH to minimize isotopic hydrogen losses. Thus, enzymes that are substantially inactivated by the required acidic buffers are of limited use in the method of the invention. However, several carboxypeptidases are enzymatically active under acid conditions, and thus are suitable for proteolysis of protein fragments under acidic conditions, *e.g.*, pH 2-3.

[0151] Most known acid-reactive proteases cleave peptides in a nonspecific manner similar to pepsin. One class of acid-reactive proteases, the carboxypeptidases, is able to generate all required subfragments of proteolytically-generated protein fragments in quantities sufficient for high resolution localization of an isotopic hydrogen label. Many carboxypeptidases are active at pH 2-3 and sequentially cleave amino acids from the carboxy terminus of protein fragments. Such enzymes include, for example, carboxypeptidases P, Y, W, and C (Breddam, *Carlsberg Res. Commun.* **51**:83-128, 1986). The need to minimize isotopic hydrogen losses precludes the use of carboxypeptidases which are inactive in acidic (pH 2.7) buffers, such as carboxypeptidases A and B.

[0152] Progressive degradation of purified isotopic hydrogen label-bearing protein fragments is preferably performed with one or more acid resistant carboxypeptidase under conditions that produce a complete set of amide-labeled subfragments, wherein each subfragment is shorter than the preceding subfragment by 1 - 5 carboxy terminal amino acids, preferably by a single carboxy-terminal amino acid. HPLC analysis of the resulting series of

subfragments allows the reliable assignment of label to a particular amide position within the parent labeled protein fragment.

[0153] In one preferred embodiment, isotopic hydrogen-labeled proteins are nonspecifically fragmented with pepsin or one or more pepsin-like proteases. The resulting labeled protein fragments are isolated by two-dimensional HPLC. These labeled protein fragments are then exhaustively subfragmented by progressive degradation with one or more acid-reactive carboxypeptidases. The resulting digests are then analyzed via RP-HPLC performed at a temperature of about 0 ° C in TFA-containing buffers (pH about 2.1). Each of the generated subfragments (typically 5 - 20) is then identified as to its structure and content of isotopic hydrogen label. The isotopic hydrogen label is thereby assigned to specific peptide amide positions.

[0154] Controlled progressive degradation from the carboxy-terminus of isotopic hydrogen labeled protein fragments with carboxypeptidases can be performed under conditions which result in the production of analytically sufficient quantities of a series of carboxy-terminal truncated subfragments, each shorter than the preceding subfragment by a single carboxy-terminal amino acid. As each carboxy-terminal amino acid of the labeled protein fragment is sequentially cleaved by the carboxypeptidase, the peptide amide nitrogen which exhibits slow hydrogen exchange under the process conditions is converted to a secondary amine which exhibits rapid hydrogen exchange. Thus any isotopic hydrogen label at that nitrogen is lost from the protein subfragment within seconds, even at acidic pH. A difference in the molar quantity of label associated with any two sequential subfragments indicates that the isotopic label is localized at the peptide bond amide between the two subfragments.

2. Location of the isotopic hydrogen label

[0155] In one preferred embodiment, synthetic peptides are produced (by standard peptide synthesis techniques) that are identical in primary amino acid sequence to each of the labeled proteolytically-generated protein fragments. The synthetic peptides may then be used in preliminary carboxypeptidase subfragmentation at a pH of about 2.7 and a temperature of about 0 ° C, and HPLC (in TFA-buffered solvents) studies to determine: 1) the optimal

conditions of proteolysis time and protease concentration which result in the production and identification of all possible carboxypeptidase products of the protein fragment under study; and 2) the HPLC elution position (mobility) of each carboxypeptidase-generated subfragment of synthetic peptide.

[0156] In one preferred aspect thereof, a set of synthetic peptides may be produced containing all possible carboxy-terminal truncated subfragments which an acid carboxypeptidase could produce upon treatment of a "parent" protein fragment. These synthetic peptides serve as HPLC mobility identity standards and enable the identification of carboxypeptidase-generated subfragments of the labeled protein fragment. Certain subfragments may be enzymatically produced by carboxypeptidase in quantities insufficient for direct amino acid analysis or sequencing. However, the quantity of the carboxypeptidase-generated subfragments is sufficient for identification by measuring HPLC mobility of such subfragments and comparing to the mobility of the synthetic peptides. Protein fragments and subfragments can be detected and quantified by standard in-line spectrophotometers (typically UV absorbance at 200 - 214 nM) at levels well below the amounts needed for amino acid analysis or gas-phase Edman sequencing.

[0157] After these preliminary studies, the proteolytically-generated HPLC-isolated, isotopically-labeled protein fragment is subfragmented with a carboxypeptidase and analyzed under the foregoing experimentally optimized conditions. The identity of each fragment is determined (by peptide sequencing or by reference to the mobility of synthetic peptide mobility marker) and the amount of isotopic hydrogen associated with each peptide subfragment is determined.

Denaturation and disulfide reduction

[0158] With some proteins, there is an absolute requirement for the employment of denaturants to effect fragmentation under quench conditions. An example of a protein with such an absolute dependency is Hen Egg White Lysozyme (HEL). In a preferred embodiment, the labeled protein is exposed, before fragmentation, to denaturing conditions compatible with slow hydrogen exchange and sufficiently strong to denature the protein enough to render it adequately susceptible to the intended proteolytic treatment. If these

denaturing conditions would also denature the protease, then, prior to proteolysis, the denatured protein is switched to less denatured conditions (still compatible with slow H-exchange) sufficiently denaturing to maintain the protein in a protease-susceptible state but substantially less harmful to the protease in question. Preferably, the initial denaturant is guanidine thiocyanate, and the less denaturing condition is obtained by dilution with guanidine hydrochloride. Guanidine hydrochloride is an effective denaturant at a concentration of about 0.05 - 4 M.

[0159] In previous studies by Englander *et al.* and others recited above, proteolytic fragmentation of labeled proteins under slowed-exchange conditions was suitably accomplished by simply shifting the protein's pH to 2.7, adding high concentrations of liquid phase pepsin, followed by (10 minute) incubation at 0 °C. With the proteins studied and reported by others to date, simply shifting pH from that of physiologic (7.0) to 2.7 was sufficient to render them sufficiently denatured as to be susceptible to pepsin proteolysis at 0 °C. Furthermore, these reported proteins, in general, did not contain disulfide bonds that interfered with effective denaturation by such (acid) pH conditions or contain disulfide bonds within portions of the protein under study with the technique.

[0160] However, in accordance with the present invention, it has been found that other proteins (for example, HEL) are negligibly denatured and are not substantially susceptible to pepsin proteolysis when continuously incubated at comparable acidic pH and depressed temperature (10 - 0 °C) for several hours. This is likely the consequence of the existence of a thermal barrier to denaturation for many proteins incubated in many denaturants; *i.e.*, denaturation of proteins at lower temperatures (10 - 0 °C), an absolute requirement for hydrogen exchange quench, is often inefficient and a slow process, incompatible with the requirement of medium resolution hydrogen exchange techniques that manipulations be performed rapidly, such that the attached label is substantially retained at functionally labeled amides of the protein.

[0161] Using the methods of the present invention, it has been discovered that such proteins become extraordinarily susceptible to pepsin proteolysis at 0 °C when they are treated with the sequential denaturation procedure described below.

[0162] While proteins are often subjected to purposeful denaturation with agents other than a pH shift prior to digestion with pepsin, this has never been done at depressed temperatures (10 - 0 °C) before, and the it has been discovered herein that while guanidine thiocyanate at the indicated concentrations is sufficient to suitably denature and render susceptible to pepsin proteolysis proteins at 10 - 0 °C, several other strong denaturants, including urea, HCl, sodium dodecyl sulfate (SDS) and guanidine HCl, were, at least when used alone, unable to adequately denature lysozyme at these low temperatures. However, the concentrations of guanidine thiocyanate required for such denaturation are incompatible with pepsin digestion; *i.e.*, they denature the pepsin enzyme before it can act on the denatured binding protein. When the guanidine thiocyanate is removed (at 10 - 0 °C) from the solution after protein denaturation has been accomplished in an attempt to overcome this inhibition of pepsin activity, the protein rapidly refolds and/or aggregates, which renders it again refractory to the proteolytic action of pepsin.

[0163] It has been discovered herein that if proteins are first denatured in about 1.5 - 4 M (preferably ≥2M) guanidine thiocyanate at 0 °C and the concentration of thiocyanate then reduced to preferably ≤0.5 M, while at the same time the guanidine ion is maintained at about 0.05 - 4 M (preferably ≥2M) (by diluting the guanidine thiocyanate- protein mixture into guanidine hydrochloride solution), the denatured protein remains in solution, remains denatured, and the enzyme pepsin remains proteolytically active against the denatured protein in this solution at 0 °C. The denatured (or denatured and reduced) protein solution is then passed over a pepsin-solid- support column, resulting in efficient and rapid fragmentation of the protein (in less than 1 minute). The fragments can be, and usually are, immediately analyzed on RP-HPLC without unnecessary contamination of the peptide mixture with the enzyme pepsin or fragments of the enzyme pepsin. Such contamination is problematic with the technique as taught by Englander *et al.*, as high concentrations of pepsin (often equal in mass to the protein under study) are employed, to force the proteolysis to occur sufficiently rapidly at 0 °C.

[0164] The stability of pepsin-agarose to this digestion buffer is such that no detectable degradation in the performance of the pepsin column employed by the methods of the present invention has occurred after being used to proteolyze more than 500 samples over 1 year. No

pepsin autodigestion takes place under these conditions. Denaturation without concomitant reduction of the binding protein may be accomplished by contacting it (at 0 - 5 °C) with a solution containing ≥2M guanidine thiocyanate (pH 2.7), followed by the addition of an equal volume of 4 M guanidine hydrochloride (pH 2.7).

[0165] Subsequent to this discovery of the extraordinary stability to denaturation of HEL under quench conditions, and the foregoing remedy, it has been found that all other proteins studied to date by methods of the present invention are susceptible, at least to a minimal degree, to pepsin proteolysis under simple quench conditions, but that the speed and extent of fragmentation can be dramatically increased by the addition of suitable concentrations of guanidine hydrochloride (GuHCl) alone, without the use of guanidine thiocyanate. There is considerable virtue in avoiding the use of thiocyanate when possible: there is a variable (often severe) aggregation and precipitation of some of the denatured protein as the thiocyanate is diluted out prior to proteolysis, greatly confounded automated sample processing.

[0166] In accordance with the present invention, it has been found that several variables behave independently in determining the speed and pattern of digestion, and that their effects are distinctive for each target protein studied. Typically, up to 30 combinations of these variables are evaluated to establish optimal fragmentation conditions for the protein under study. These independent variables include the type of denaturant (*e.g.*, GuSCN versus GuHCl); its concentration preferably (0.05 - 4M); the time the denaturant is allowed to act on the protein prior to fragmentation (preferably 0 to 3 minutes); the type(s) of endoproteinases employed; and the time allowed for digestion (preferably 20 seconds to 2 minutes). For most proteins studied, GuHCl, at a concentration of 0.5M and 30 seconds fragmentation on a pepsin column as above is near-optimal, though more extensive tuning will likely improve the fragmentation map.

[0167] It is to be emphasized that the speed of generation (typically in 30 seconds) and the yield and extent of the highly overlapping fragmentation seen using the high resolution hydrogen exchange methods presented herein is unprecedented in the previously disclosed art, and was unanticipated until these recent results. There was no expectation that the art of modulating endopeptidase activity-both in terms of producing the needed varied

fragmentation and yield could be enhanced enough to be useful by itself for high resolution localization of label. Heavy hydrogen label is quickly lost from proteolytic fragments during analysis, even under quench conditions: thus, all steps of analysis should be performed as quickly as possible, including protease digestion. The methods developed and available prior to 1997 required pepsin degradation durations that were already at the upper limits of acceptable times (approximately 10 minutes). For example in U.S. Patent No. 6,291,189, it is stated that : "In a preferred embodiment, pepsin is used, preferably at a concentration of 10 mg/ml pepsin at 0° C, pH 2.7 for 5-30 minutes, preferably 10 minutes." It was therefore unanticipated that more extensive digestions could be obtained with pepsin with or without other endoproteinases given the time constraints of amide hydrogen exchange study.

[0168] Accordingly, the methods of the present invention analyze endopeptidase fragments that are generated by cleaving the labeled protein with an endopeptidase selected from the group consisting of a serine endopeptidase, a cysteine endopeptidase, an aspartic endopeptidase, a metalloendopeptidase, and a threonine endopeptidase (a classification of endopeptidases by catalytic type is available on the world wide web at the URL "chem.qmul.ac.uk/iubmb/enzyme/EC34"; by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology). Presently preferred endopeptidases include pepsin, newlase and acid tolerant Aspergillus proteases such as Aspergillus protease XIII. It has further been found that the fragmentation patterns resulting from simultaneous, and/or sequential proteolysis by combinations of these enzymes are additive in their effect on fragmentation. Therefore, more than one endopeptidase may be used in combination. Optimally, endopeptidase fragments are generated at a pH of about 1.8 - 3.4, preferably 2-3, more preferably in the range of about 2.1 - 2.3 or 2.5 - 3.0.

[0169] In preferred embodiments, the endopeptidase may be coupled to a perfusive support material to facilitate manipulation of digestions, as an alternative to liquid phase digestions. This allows the reuse of endopeptidase materials and separates the enzyme from the fragments for further analysis. Exemplary perfusive support matrices include Poros 20 media, wherein digestion of the labeled protein is accomplished by contacting a solution of the labeled protein with said matrix, followed by elution of generated fragments from the matrix. With the use of the solid support, sample digestion under slowed exchange

conditions can be performed that results in no detectable endoproteinase autodigestive fragments being released into the digestion product, *i.e.*, the population of labeled protein fragments. Furthermore, the endoproteinases remain fully active and available for subsequent repeated use as a digestive medium for additional samples.

[0170] In accordance with the present invention, it has been discovered herein that the judicious admixture of denaturants with substrate protein results in the ability to greatly promote and "tune" substrate fragmentation. Unfortunately, these same denaturants retard and/or inhibit the activity of the enzymes unless denaturants are partially removed prior to proteolysis. However their removal allows the substrates to re-fold, negating the benefit of the denaturant. Gradual manual dilution of the substrate-enzyme-denaturant mixture allowed an initially slow proteolysis to proceed. With subsequent dilution, partially degraded substrate is unable to refold; and because of denaturant dilution, protease activity increases, further fragmenting the initial large substrate fragments. Success in this method required multiple manual additions of reagents, denaturants, and timed addition of diluents, all very labor intensive. The improved methods of the present invention use solid-state enzymes on perfusive supports and column chromatography, enabling samples to be applied to the column already mixed with denaturant, and the necessary dilution of denaturant automatically occurs as the substrate slug passes down the column, now progressively diluted with the fluid in the column void volume as proteolysis proceeds. This results in tremendous labor savings, and is readily automated. There is thus an unanticipated ease and simplification of use of the necessary denaturants when solid phase proteases are employed.

[0171] A variety of acid-reactive endoproteinases can be covalently coupled to any of a number of available support matrices including, for example, cross-linked dextran, cross-linked agarose, as well as more specialized supports suitable for modern HPLC chromatography, preferably the Poros line of perfusive support materials supplied by Perceptive Biosystems, such as "20-AL" and the like. These latter supports are particularly advantageous for invention methods as they allow rapid interaction of substrate with bound peptidases. The coupling of endoproteinases to matrices can be achieved by any of a number of well-known chemistries capable of effecting such couplings, including, for example, aldehyde-mediated (sodium cyanoborohydride-stabilized Schiff base), carbodiimide, and

cyanogen bromide-activated couplings. Conditions, including pH, conducive to the continued stability of particular peptidases may optionally be employed, and could readily be implemented by one of skill in the art.

[0172] An exemplary preparation of coupled endopeptidase is as follows. The endopeptidase is obtained as a lyophilized powder, reconstituted with distilled water, and dialyzed against a coupling buffer containing 50 mM citrate (pH 4.4). The peptidase is then coupled to Perceptive Biosystems Poros media 20-AL following the manufacturer's recommended coupling procedures, including "salting out" with high sodium sulfate concentrations. Couplings can be performed at a ratio of 5 to 30 mg of peptidase per ml of settled 20-AL matrix, preferably 30 mg/ml. The coupled matrix can then be stored in the presence of sodium azide to minimize bacterial contamination.

[0173] While any of a number of batchwise or column chromatographic approaches might be employed to effect matrix-bound endopeptidase digestion of labeled protein under slowed exchange conditions, the following approach has been found to work well and to be preferable. A stainless steel column (length 2 cm, width 2.2 mm, internal volume approximately 66 microliters) was packed with endoproteinase-derivatized 20-AL support coupled with protein at 30 mg/ml) and flow established with a solvent consisting of 0.5% formic acid (for pepsin, newlase, or Aspergillus protease XIII), said column being operated at 0 °C. Care must be taken to employ buffers with a pH compatible with rapid peptidase action: buffers with a pH of 2.7- 3.0 (room temperature measurement) work well. An aliquot of labeled protein to be fragmented was contacted with the column matrix typically in a volume of 10 - 300 microliters, preferably 100 microliters, and the sample allowed to reside on the column for a time determined (by preliminary titration studies) to result in the desired degree of fragmentation. It has been surprisingly found herein that digestion times of 13 seconds to 5 minutes, preferably less than a minute, more preferably, less than 40 seconds to be optimal. Prior knowledge of endopeptidase digestion suggested that digestion times of greater than 10 minutes would be required to produce sufficient fragmentation. The sample was then flushed from the column onto either an analytical reverse phase HPLC column for subsequent separation and analysis of the peptide fragments, or directly without additional purification or chromatography onto a mass spectrometer for analysis. During this analysis

period, the column is flushed (with the effluent going to waste) with an excess of solvent to remove any peptide or subfragments which nonspecifically adhere or are otherwise retained in the matrix, thereby preparing the column for a repeated use. Such washing buffers can be any of a wide variety of buffers including the buffers used for digestion. The column-washing step (between each sample digestion) is preferable but not absolutely required for success.

[0174] In an additional embodiment, a column containing one of these solid state proteases can be used to further digest peptides on-line as they each independently exit the reversed phase (RP) HPLC column during gradient elution. This approach has the considerable advantage of producing a much less complex mixture of peptides to analyze than when two enzymes act on the substrate before RP-HPLC. To use these enzymes in this post-chromatography manner, it may be useful to reduce the acetonitrile concentration in the effluent stream prior to passage over the protease column, as acetonitrile can reversibly (and irreversibly) inhibit these enzymes.

[0175] In addition, disulfide bonds, if present in the protein to be digested, can also interfere with analysis. Disulfide bonds can hold the protein in a folded state where only a relatively small number of peptide bonds are exposed to proteolytic attack. Even if some peptide bonds are cleaved, failing to disrupt the disulfide bonds would reduce resolution of the peptide fragments still joined to each other by the disulfide bond; instead of being separated, they would remain together. This would reduce the resolution by at least a factor of two (possibly more, depending on the relationship of disulfide bond topology to peptide cleavage sites).

[0176] In one embodiment, water soluble phosphines, for example, Tris (2-carboxyethyl) phosphine (TCEP) may be used to disrupt a protein's disulfide bonds under "slow hydrogen exchange" conditions. This allows much more effective fragmentation of large proteins which contain disulfide bonds without causing label to be lost from the protein or its proteolytic fragments (as would be the case with conventional disulfide reduction techniques which must be performed at pHs which are very unfavorable for preservation of label).

[0177] High resolution localization of label-bearing amides with the use of endoproteinases requires the proteolytic generation of numerous sequence-overlapped fragments under conditions which allow the label to remain in place (*e.g.*, 0 °C, pH 2.2). The ability of any protease to fragment a protein or peptide is limited by the accessibility of the protease to susceptible peptide bonds. While denaturants such as acidic pH, urea, detergents, and organic co-solvents can partially denature proteins and expose many otherwise structurally shielded peptide bonds, pre-existing disulfide bonds within a protein can prevent sufficient denaturation with these agents alone. In conventional protein structural studies, disulfides are usually cleaved by reduction with 2-mercaptoethanol, dithiothreitol, and other reductants which unfortunately require a pH greater than 6 and elevated temperature for sufficient activity, and are therefore not useful for the reduction of disulfides at pH 2.7 or below. For this reason, the hydrogen exchange art has not attempted any form of disulfide bond disruption, has for the most part been restricted to the study of proteins without intrinsic disulfide bonds, and has accepted the low resolution achievable without disulfide bond disruption.

[0178] It has been recognized and demonstrated herein that acid-reactive phosphines such as Tris (2-carboxyethyl) phosphine (TCEP) can be used to disrupt disulfides under the acidic pH and low temperature constraints required for hydrogen exchange analysis. These manipulations disrupt these associations and at the same time continue to produce a markedly slowed proton exchange rate for peptide amide protons.

[0179] Denatured (with or without reduction) labeled protein is then passed over a column composed of insoluble (solid state) pepsin, whereby during the course of the passage of such denatured or denatured and reduced binding protein through the column, it is substantially completely fragmented by the pepsin to peptides of size range 2-20 amino acids at 0 °C and at pH 2.7. The effluent from this column (containing proteolytically-generated fragments of labeled protein) is directly and immediately applied to the chromatographic procedure employed to separate and analyze protein fragments, preferably analytical reverse-phase HPLC chromatography and/or mass spectrometry.

[0180] In preferred embodiments, proteins containing disulfide bonds may be first physically attached to solid support matrices, and then contacted with solutions containing

TCEP at acidic pH and low temperature for more rapid reactions than are possible in solution. In this preferred embodiment, with all steps performed at 5 - 0 °C, preferably 0 °C, the protein in aqueous solution, with or without prior denaturation and under a wide variety of pH conditions (pH 2.0 - 9.0) is first contacted with a particulate silica-based reverse-phase support material or matrix typically used to pack HPLC columns, including C4 and C18 reversed phase silica supports, thereby attaching the protein to the surface of such material. Unbound binding protein may then optionally be washed off the support matrix with typical aqueous HPLC solvents, (0.1 % trifluoroacetic acid, (TFA) or 0.1-0.5 % formic acid in water, buffer A). An aliquot of a substantially aqueous buffer containing TCEP at a pH between 2.5 and 3.5, preferably 2.7 is then contacted with the protein that is attached to the support material and allowed to incubate with the attached protein near 0 °C and preferably for short periods of time (0.5-20 minutes, preferably 5 minutes) and then the TCEP-containing buffer removed from the support matrix by washing with buffer A , followed by elution of the reduced binding protein from the support matrix by contacting the support with eluting agents capable of disrupting the support- protein binding interaction, but also compatible with continued slow hydrogen exchange (pH 2.0-3.5; temperature 0 - 5 °C).

[0181] An example of this preferred embodiment to achieve disulfide reduction prior to pepsin fragmentation is as follows. Labeled protein is applied to a reverse phase silica- based C18 HPLC support matrix in a column (for example, Vydac silica- based C18, catalog #218TP54, or Phenominex silica- based C18 Jupiter 00B4053-B-J) that has been pre-equilibrated with HPLC solvent A (0.1% TFA or 0.1 - 0.5% formic acid at 0 - 5 °C. After substantial binding of the lysozyme has occurred (usually within seconds), additional buffer A is passed through the column to remove small quantities of unattached binding protein. A solution containing TCEP (50 - 200 micrometers of TCEP (0.05 - 2.0 M in water at a pH of 2.5-3.5, preferably 3.0) is then applied to the column in a manner that results in its saturation of the portion of the column to which the binding protein has been previously attached. Flow of solvent on the support is then stopped to allow incubation of the TCEP solution with the support matrix-attached binding protein. At the end of this incubation time (variously 0.5 minutes - 20 minutes, preferably 5 minutes) flow of solvent A is resumed, resulting in the clearance and washing of the TCEP solution from the support matrix. This is followed by application of an amount of solvent B (20% water, 80% acetonitrile, 0.1% TFA) sufficient to

release the binding protein from the support (typically 30-50% solvent B in solvent A). This eluted and reduced protein is then passed over a pepsin column to effect its fragmentation under slowed exchange conditions. The protein fragments resulting from the action of the pepsin column on the reduced protein are then contacted with another analytical HPLC column, preferably a reverse phase HPLC support, and the fragments sequentially eluted from the support with a gradient of solvent B in solvent A.

[0182] An example of an alternative preferred embodiment to achieve disulfide reduction after pepsin fragmentation is as follows. This alternative approach is to first denature the protein under slow exchange conditions, pass it over a pepsin column to effect fragmentation, apply the resulting fragments to a HPLC support matrix, effect reduction of the support-bound peptide fragments by contacting them with the aforementioned TCEP solution, followed by sufficient incubation at 0 °C, finally followed by elution of the reduced fragments from the column with increasing concentrations of solvent B. The advantage of this second alternative method is that an entire HPLC support matrix attachment-detachment step is avoided, resulting in a simplification of the manipulations and equipment required for the procedure, as well as savings in elapsed time. This approach is not probable when a particular protein requires substantial prior reduction of disulfides to become substantially susceptible to the digestive actions of pepsin. Certain proteins are sufficiently stabilized by their contained disulfide bonds that they may not become substantially susceptible to pepsin even in the presence of strong denaturants. In such cases it will be preferable to apply the first method of reduction (above), where the protein is first reduced "on column", eluted, fragmented on the pepsin column, and the fragments then optionally applied to an additional column matrix to effect separation from each other.

[0183] Additionally, it has been found herein that the simultaneous use of denaturants and reductants (TCEP) results in synergistic enhancement of both protein denaturation and reduction, not seen when employed separately, or even sequentially.

Deconvolution of endopeptidase-generated fragments in methods employing improved proteolysis fragmentation

[0184] Mass spectroscopy has become a standard technology by which the amino acid sequence of proteolytically generated peptides can be rapidly determined. It is commonly used to study peptides which contain amino acids which have been deuterated at carbon-hydrogen positions, and thereby determine the precise location of the deuterated amino acid within the peptide's primary sequence. This is possible because mass spectroscopic techniques can detect the slight increase in a particular amino acid's molecular weight due to the heavier mass of deuterium. McCloskey (*Meth. Enzymol.* **193**:329-338, 1990) discloses use of deuterium exchange of proteins to study conformational changes by mass spectrometry. The methods of the present invention include measuring the mass of endopeptidase-generated fragments to determine the presence or absence, and/or the quantity of deuterium on the endopeptidase-generated fragments. Preferably, mass spectrometry is used for mass determination of these peptide fragments. This allows determination of the quantity of labeled peptide amides on any peptide fragment.

[0185] According to the methods of the present invention, proteolytically generated fragments of protein functionally labeled with deuterium may be identified, isolated, and then subjected to mass spectroscopy under conditions in which the deuterium remains in place on the functionally labeled peptide amides. Standard peptide sequence analysis mass spectroscopy can be performed under conditions which minimize peptide amide proton exchange: samples can be maintained at 4 °C to 0 °C with the use of a refrigerated sample introduction probe; samples can be introduced in buffers which range in pH between 1 and 3; and analyses are completed in a matter of minutes. MS ions may be made by MALDI (matrix-assisted laser desorption ionization) electrospray, fast atom bombardment (FAB), *etc.* Fragments are separated by mass by, *e.g.*, magnetic sector, quadropole, ion cyclotron, or time-of-flight methods. For MS methods generally, see Siuzdak, G., Mass Spectrometry for Biotechnology (Academic Press 1996).

[0186] Once the endopeptidase fragmentation data is acquired on functionally deuterated protein, it is then deconvoluted to determine the position of labeled peptide amides in an amino acid specific manner. In general, the term "deconvoluted" as used herein refers to the mapping of deuterium quantity and location information obtained from the fragmentation data onto the amino acid sequence of the labeled protein to ascertain the location of labeled

peptide amides, and optionally their rates of exchange. Deconvolution may comprise comparing the quantity and/or rate of exchange of isotope(s) on a plurality of endopeptidase-generated fragments with the quantity and rate of exchange of isotope(s) on at least one other endopeptidase fragment in the population of fragments generated, wherein said quantities are corrected for back-exchange in an amino acid sequence-specific manner. Labeled peptide amides can optionally be localized in an amino acid sequence-specific manner by measuring rates of off-exchange of functionally attached label under quenched conditions. The determination of the quantity and rate of exchange of peptide amide hydrogen(s) may be carried out contemporaneously with the generation of the population of endopeptidase-generated fragments.

[0187] Although several alternative methods for effecting such deconvolution may be available, at least one useful method has been implemented and demonstrated herein. As presented in the Examples herein, a protein construct composed of a two repeat segment (16-17) of chicken brain spectrin was on- exchanged in deuterated buffer for varying times (10 to 100,000 seconds, at 22 °C). Samples were then exchange-quenched, in 0.5 M GuHCl, pH 2.2, and processed. The deuterium content of the 113 useful peptides resulting from such fragmentation was determined from the raw MS data, with corrections for back-exchange made employing the inexact "peptide average" method of Zhang and Smith (Zhang *et al., Prot. Sci.* **2**:522-531, 1993).

[0188] Plots of deuterium buildup versus time were constructed for each peptide, and the number of amides exchanging in arbitrary "fast, medium and slow" classes determined for each peptide. An initial map of rates versus amino acid sequence was then constructed from this information, employing a strategy in which "pieces" (fragments) with uniform rate class were first placed in register, and subsequent placement of more complexly patterned pieces (two rate classes then three) performed in a manner that required that the several rate classes in these peptides be reconciled vertically to conform with the placement of the preceding pieces. The average rate class at each amide position was then determined and used to construct the initial map. Unmeasureable amide hydrogens (approximately 10% of the total amides in the 113 fragments, unmeasured either because of errors incurred because of the approximate (average) back-exchange calculation method employed, or because the very

slowest exchanging amides were not measured in this experiment) were then fit to the provisional map in a manner that minimized deviation from said map, and a final map constructed by averaging this final placement of "pieces".

[0189] The choice of three rate classes was arbitrary, and done to simplify the "piece placement" work, which was done manually in this example. Assignment of amides in each peptide to each of 9 rate classes (9 time points were employed in this experiment) would considerably improve the resolution of the deconvolution, but is conveniently performed by automated (computational) means, and with incorporation of more precise back-exchange corrections as discussed below. Further fragmentation of this protein construct with pepsin plus Fungal protease XIII has resulted in a 50 % increase in the number of spectrin fragments, which will preferably be deconvoluted through linear programming-mediated approaches.

[0190] The essential attributes of a preferred deconvolution algorithm for such high density, overlapping endopeptidase fragment data include that: (i) it takes as inputs the measurements of the quantity of label on the numerous overlapping endopeptidase-generated fragments correlated with their amino acid (aa) sequence; (ii) it more precisely corrects for back-exchange (that is, label lost subsequent to initiation of quench, during the analysis step) than the presently employed method that calculates an average correction factor for all amides in a peptide (Zhang *et al., Prot. Sci.* 2:522-531, 1993) and instead employs a correction that is sub-site-specific (specific for 1-5 contiguous amides, depending on the resolving power of the aggregate endopeptidase fragments available). This can be done both computationally (by reference to the Bai/Englander-algorithm; ; Bai *et al., Proteins: Struct. Funct. Genet.* 17:75-86, 1993) or alternatively experimentally by measuring back exchange, under quench conditions, of the substantially random coil fragments resulting from identical endoproteolysis of a fully (equilibrium) deuterated sample of the protein in a manner that allows the rate(s) of loss of deuterium to be measured over time for each resolvable sequence region. Either approach affords precise calculation of the label lost through back exchange from each peptide, and, by comparison, that lost in each amino acid segment that differs between amino acid sequence-overlapping peptides. Corrections for these losses are made for each peptide/amino acid overlap difference value; (iii) it compares the (corrected) label

content of each peptide with the label content of all peptides with which it (or immediately adjoining peptides) share any part of amino acid sequence, said comparisons being performed in a manner which allows differences in label content to be assigned to regions of amino acid sequence difference, with the preferred algorithm seeking to fit deuterium location and quantity at each location in a manner that optimizes agreement between results obtained from the plurality of fragments; and (iv) it optionally makes use of measurements of off-exchange rates of label on quenched fragments, which, by reference to the above noted site- specific rate (under quench conditions) prediction or empirical determination from endoproteinase fragmentation data of equilibrium- deuterated protein) can be employed to further sublocalize label at regions unresolved by analysis of fragments alone at one quench condition duration.

Automation of hydrogen exchange analyses

[0191] The high resolution hydrogen exchange methods of the present invention may be performed using an automated procedure. Automation may be employed to perform isotope-exchange labeling of proteins as well as subsequent proteolysis and MS-based localization procedures. The use of such automation allows one to manipulate proteolysis conditions under quench conditions, largely by employing solid-state chemistries as described above. The following discussion refers to modules as designated in the exemplary deuterium exchange-mass spectrometry (DXMS) apparatus. The fluidics of the DXMS apparatus contains a number of pumps, high pressure switching valves and electric actuators, along with connecting tubing, mixing tees, and one way flow check valves and that direct the admixture of reagents and their flow over the several small stainless steel columns containing a variety of proteins and enzymes coupled to perfusive (Poros 20) support material.

[0192] While there is a standard configuration of these various components, the pattern of the several elements can be quickly changed to suit particular experimental requirements. DXMS fluidics contains a "cryogenic autosampler" module (A), a "functional deuteration" or sample preparation module (B) used for automated batched processing of manually prepared samples, and a "endopeptidase proteolysis" module (C). Precise temperature control is achieved by enclosing the valves, columns, and connecting plumbing of modules A, B, and C in a high thermal-capacity refrigerator kept at about 3.8 ° C (the freezing point of deuterated

water), and components that have no contact with pure deuterated water are immersed in melting (regular) ice.

[0193] Module A, the "cryogenic autosampler" allows a sample set (in the range of about 10-50 samples) to be prepared manually in autosampler vials, quenched, denatured, and samples frozen at -80 °C, conditions under which loss of deuterium label in the prepared samples is negligible over weeks. This allows a large number of deuterated samples to be manually prepared, and then stored away for subsequent progressive proteolysis. This capability also allows samples to be manually prepared at a distant site, and then shipped frozen to the DXMS facility for later automated analysis. This module contains a highly modified Spectraphysics AS3000® autosampler, partially under external PC control, in which the standard pre-injection sample preparation features of the autosampler are used to heat and melt a frozen sample rapidly and under precise temperature control. Under computerized control, the autosampler's mechanical arm lifts the desired sample from the -80 °C sample well, and places it in the autosampler heater/mixer/vortexer which rapidly melts the sample at 0 - 5 °C. The liquified sample is then automatically injected onto the HPLC column.

[0194] Optional modifications to a such a standard autosampler may include: modification of the sample basin to provide an insulated area in which dry ice can be placed, resulting in chilling of the remaining areas of the sample rack to -50 to -80 °C; placement of the autosampler within a 0 - 5 °C refrigerator, and "stand-off" placement of the sample preparation and sample injection syringe assemblies of the autosampler outside the refrigerator, but with otherwise nominal plumbing and electrical connection to the autosampler. An external personal computer (PC) (running Procom, and a dedicated Procrom script "Asset1"), delivers certain settings to firmware within the autosampler, allowing: (i) a much shortened subsequent post-melting dwell time of samples in the chilled basin, avoiding re-freezing of sample prior to injection; and (ii) allowing its heater/mixer to regulate desired temperatures when they are less than the default minimum temperature of 30°C. The "sample preparation" module (B), automatically performs the "functional deuteration" or sample preparation manipulations, quench, and denaturation in large part through use of the solid-state inventions as described earlier herein, for example, using a protein conjugated to

solid phase beads. Several components of this module will benefit from the microfluidics inventions also described earlier. Typically, deuterated samples are manually prepared (both at 0 °C, and at room temperature) by diluting 1 μL of protein stock solution with 19 μL of deuterated buffer (150 mM NaCl, 10 mM HEPES, pD 7.4), followed by "on-exchange" incubation for varying times (10 sec, 30 sec, 100 sec, 300 sec, 1000 sec, 3000 sec) prior to quenching in 30 μL of 0.5% formic acid, 2 M GuHCl, 0 °C. These functionally deuterated samples are then subjected to DXMS processing, along with control samples of undeuterated and fully deuterated protein (incubated in 0.5% formic acid in 95% D_2O for 24 hours at room temperature). The centroids of probe peptide isotopic envelopes are then measured using appropriate software. In order to obtain the deuteration levels of each peptide corrected to the values after "on-exchange" incubation, but before DXMS analysis, the corrections for back-exchange are made employing the methods of Zhang and Smith as previously described.

[0195] Regardless of the manner of sample preparation, quenched samples are then automatically directed to the "proteolysis" module (for methods employing progressive proteolysis fragmentation), or alternatively the "endopeptidase proteolysis" module (C) (for methods employing improved proteolysis fragmentation), in which proteolysis is accomplished using a battery of solid-state protease columns, variously pepsin, fungal protease XIII, newlase, *etc.* as desired, with the resulting peptide fragments being collected on a small reversed-phase HPLC column, with or without the use of a small c18 collecting pre-column. This column(s) is then acetonitrile gradient-eluted, with optional additional post-LC on-line proteolysis. The effluent is then directed to the electrospray head of the mass spectrometer (a Finnegan ion trap or a Micromass Q-TOF) which protrudes into a hole drilled in the side of the refrigerator. Several components of this module lend themselves to microfluidic devices as described earlier.

[0196] In a preferred embodiment, the proteolysis module contains four high pressure valves (Rheodyne 7010); with valve 1 bearing a 100 μL sample loop; valve 2 bearing a column (66 μL bed volume) packed with porcine pepsin coupled to perfusive HPLC support material (Upchurch Scientific 2 mm x 2 cm analytical guard column; catalog no. C.130B; porcine pepsin, Sigma catalog no. p6887, coupled to Poros 20 AL media at 40 mg/mL, in 50 mM sodium citrate, pH 4.5, and packed at 9 mL/min according to manufacturer's

instructions); valve 3 bearing a C18 microbore (1 mm x 5 cm) reversed phase HPLC column (Vydac catalog no. 218MS5105), and valve 4 connected to the electrospray head of a mass spectrometer. Inline filters (0.05 um, Upchurch catalog no. A.430) are placed on each side of the pepsin column, and just before the C18 column (Vydac prefilter, catalog no. CPF 10) to minimize column fouling and carryover from aggregated material.

[0197] In this configuration, four HPLC pumps (Shimadzu LC-10AD, operated by a Shimadzu SCL-10A pump controller) supplied solvents to the valves; with pumps C and D providing 0.05 % aqueous TFA to valve 1 and valve 2 respectively; pumps A (0.05% aqueous TFA) and B (80% acetonitrile, 20% water, 0.01% TFA) are connected through a microvolume mixing tee (Upchurch catalog no. P.775) to provide valve 3 with the C18 column-eluting gradient. All valves are connected to Two-Position Electrical Actuators (Tar Designs Inc.).

[0198] A typical sample is processed as follows: a 20 μL of hydrogen exchanged protein solution is quenched by shifting to pH 2.2 - 2.5, 0 °C with a 30 μL of quenching stock solution chilled on ice. The quenched solution is immediately pulled into the sample loading loop of valve 1, and then the computer program (see below) started. Pump C flow (0.05% TFA at 200 uL/min) pushes the sample out of injection loop onto the C18 HPLC column via the solid-state pepsin column at valve 2 (digestion duration of about 26 seconds). After two minutes of pump C solvent flow, the C18 column is gradient-eluted by pumps A and B (linear gradient from 10 to 50 % B over 10 minutes; 50 uL/min; pumps A, 0.05% TFA; pump B, 80% acetonitrile, 20% water, 0.01% TFA), with effluent directed to the mass spectrometer. During data acquisition, pump D (aqueous 0.05% TFA 1 mL/min, 10 minutes) back-flushes the pepsin column to remove retained digestion products.

[0199] The timing and sequence of operation of the foregoing DXMS fluidics may be controlled by a personal computer running a highly flexible program in which sequential commands to targeted solid state relays can be specified, as well as variably timed delays between commands. Certain command lines may access an array matrix of on- and off-exchange times, and the entire sequence of commands may be set to recycle, accessing a different element of the array with each cycle executed. Certain command lines may be set to receive "go" input signals from peripherals, to allow for peripheral-control of cycle

progression. A library of command sequences may be prepared, as well as a library of on/off time arrays. An exemplary protein machine program can be configured to execute a supersequence of command sequence-array pairs.

[0200] An exemplary protein machine program (written in LabView I, National Instruments, Inc) controls the state(s) of a panel of solid-state relays on backplanes (SC-206X series of optically isolated and electromechanical relay boards, National Instruments, Inc.) with interface provided by digital input/output boards (model no. PCI-DIO-96 and PCI-6503, with NI- DAQ software, all from National Instruments, Inc.). The solid-state relays in turn exert control (contact closure or TTL) over pumps, valve actuators, and mass spectrometer data acquisition. Each of these peripherals are in turn locally programmed to perform appropriate autonomous operations when triggered, and then to return to their initial conditions. The autosampler and HPLC column pump controller are independently configured to deliver a "proceed through delay" command to the Digital I/O board as to insure synchronization between their subroutines and the overall command sequence.

[0201] In order to optimize or "tune" endopeptidase proteolysis, preliminary proteolytic "tuning" studies are performed to establish the fragmentation conditions (compatible with slowed exchange) optimal for peptide generation from the target polypeptide. Two major parameters that are often optimized are the concentration of GuHCl in quenching buffer and the pump C flow rate over the pepsin column. Typically, a 1 ml stock solution of protein (10 mg/ml, pH 7.0) is diluted with 19 mL of water and then quenched with 30 mL of 0.5% formic acid containing various concentrations of GuHCl (0 - 6.4 M). The quenched sample is then pulled into the sample loading loop, and the DXMS program sequence triggered immediately after sample loading. The flow over the pepsin column is varied (100 uL/min – 300 uL/min) to adjust the duration of proteolytic digestion.

[0202] In order to quickly identify pepsin generated peptides for each digestion condition employed, spectral data is preferably acquired in particular modes, for example designated herein as "triple play" and "standard double play" modes, which have been empirically tuned to optimize the number of different parent ions upon which MS2 is performed. This data is then analyzed by appropriate software.

[0203] Triple play contains three sequentially executed scan events; first scan, MS1 across 200-2000 *m/z*; second scan, selective high resolution "zoom scan" on most prevalent peptide ion in preceding MS1 scan, with dynamic exclusion of parents previously selected; and third scan, MS2 on the same parent ion as the preceding zoom scan. The triple play data set or double play data set is then analyzed employing the Sequest software program (Finnigan Inc.) set to interrogate a library consisting solely of the amino acid sequence of the protein of interest to identify the sequence of the dynamically selected parent peptide ions.

[0204] This tentative peptide identification is verified by visual confirmation of the parent ion charge state presumed by the Sequest program for each peptide sequence assignment it made. This set of peptides is then further examined to determine if the "quality" of the measured isotopic envelope of peptides was sufficient (adequate ion statistics, absence of peptides with overlapping *m/z*) to allow accurate measurement of the geometric centroid of isotopic envelopes on deuterated samples.

[0205] According to an additional aspect of the present invention, it may be useful to perform *in vivo* analysis of a polypeptide of interest, for example, *in situ* analysis of a protein, or protein-binding partner interactions. In such applications, the protein, while present in its native environment as a component of an intact living cell, or as a component of a cellular secretion such as blood plasma, is on-exchanged by incubating cells or plasma in physiologic buffers supplemented with tritiated or deuterated water. Optionally, the binding partner is then added, allowed to complex to the cell or plasma-associated protein, and then off-exchange initiated by returning the cell or plasma to physiologic conditions free of tritiated or deuterated water. During the off-exchange period (hours to days) the formed protein or complex is isolated from the cell or plasma by any purification procedure which allows the protein or complex to remain continuously intact.

[0206] According to an additional aspect of the present invention , the on- exchanged cell, plasma or other mixture prepared as above can be shifted to exchange "quench" conditions, and then the protein of interest purified under continued quench conditions, employing variously reverse phase chromatography with or without cation- exchange chromatography, with or without prior fragmentation with proteases, followed by mass spectroscopic anqalysis again under continued quench conditions. Alternatively, a desired

species can be isolated from the quench mixture employing any affinity method that operates (binding interactions occur) under quench conditions, for example, trough use of binding pairs known to operate in acid physiologic interactions, including pepsin- pepstatin interaction, cobalamin, or transcobalamin and vitamin B12, and the like. Additionally monoclonal antibodies can be prepared employing phage display techniques, in which antibodies can be produced that bind to protein epitopes under quench conditions. Such antibodies can be prepared to specific proteins one desires to purify from the above hydrogen-exchanged quench mix, or alternatively, can be prepared to generic protein sequences that can be expressed as fusion proteins with the protein of interest in the quenched mix. These include His- six tag sequences, FLAG sequences, as well as green fluorescent protein and other often used fusion sequences. For a particular affinity binding pair, the practitioner would engineer one of the binding partners into the target protein by any of a variety of recombinant DNA and express the fusion protein in an expression system using any of a variety of gene transfer and expression techniques, and employ the other member of the binding pair for affinity capture of the binding partner (and its attached target protein), for example by solid- state affinity chromatography, or magnetic bead affinity capture techniques. Desired proteins can then be eluted from such acid- stable binding supports by a number of methods including chaotropic agents, addition of excess of binding partner (without fusion partner) and the like. This analytic method is especially appropriate for proteins which lose substantial activity as a result of purification, as binding sites may be labeled prior to purification.

[0207] According to further aspects of the present invention, after determining the binding sites of a binding protein or a binding partner, by the present methods (alone or in conjunction with other methods), the information may be exploited in the design of new diagnostic or therapeutic agents. Such agents may be fragments corresponding essentially to said binding sites (with suitable linkers to hold them in the proper spatial relationship if the binding site is discontinuous), or to peptidyl or non-peptidyl analogues thereof with similar or improved binding properties. Alternatively, they may be molecules designed to bind to said binding sites, which may, if desired, correspond to the paratope of the binding partner.

[0208] The diagnostic agents may further comprise a suitable label or support. The therapeutic agents may further comprise a carrier that enhances delivery or other improves the therapeutic effect. The agents may present one or more epitopes, which may be the same or different, and which may correspond to epitopes of the same or different binding proteins or binding partners.

DXMS Methodology in Practice

[0209] Detailed descriptions of the several enhancements that constitute DXMS are presented in four articles that have recently published (Hamuro, et al. J. Mol. Biol. 323:871-881 2002, Hamuro, et al. J. Mol. Biol. 4:703- 714 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003), and four US patents issued to the PI (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Hamuro, et al. J. Mol. Biol. 4:703- 714 2002, Woods-Jr. U.S. Patent No. 5,658,739 (1997), Woods-Jr. U.S. Patent No. 6,291,189 (2001), Woods-Jr. U.S. Patent No. 6,331,400 (2001), Burns, et al. Protein Science 11:185 2002). The technique has an initial exchange-labeling step performed under entirely physiologic conditions of pH, ionic strength, and buffer salts and a subsequent localization step performed under non-native, exchange- "quench" conditions. The labeling is performed by simply adding deuterated water to a solution of the protein. During this on-exchange incubation, deuterium exchanges onto the several amides of the protein. As labeling progresses, aliquots are exchange- "quenched" by shifting the protein to conditions (low pH, and temperature) that dramatically slow the rate of exchange, effectively locking in place the attached deuterium. Undesired "back-exchange", or loss of label after establishment of sample quench, can be essentially halted by holding samples at very low temperatures (-80°C) until they are melted (at 0° C) and further processed as below. This process has been automated with the development of a cryogenic autosampler within our DXMS apparatus (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Woods-Jr. U.S. Patent No. 5,658,739 (1997), Woods-Jr. U.S. Patent No. 6,291,189 (2001), Woods-Jr. U.S. Patent No. 6,331,400 (2001)).

[0210] In the second step, the amino acid sequence location and amount of attached deuterium is determined. Under "quench" conditions, the protein sample is (automatically)

first optionally denatured, optionally disulfide- reduced, and then proteolyzed by solid-phase pepsin into overlapping fragments of ~ 3-15 amino acids in size. It is to be emphasized that this is high-throughput, exhaustive (not limited) proteolysis, with typical digestion times being of the order of 20 seconds. The digests are then subjected to rapid high performance liquid chromatography (HPLC) separation (5-10 minute gradients), and directly analyzed by electrospray-ion trap or time of flight (TOF) mass spectrometry performed under conditions adapted to amide hydrogen exchange studies.

[0211] The extent of pepsin digestion is finely tuned with the goal of generating multiple overlapping fragments of the protein. When even finer fragmentation is achieved with additional acid-reactive proteinases. Fragmentation is followed by rapid sequence identification, performed first with undeuterated protein under quench conditions, followed by assessment of deuterium label bound to each fragment generated from deuterated protein samples. The differences in deuterium content between peptides with overlapping sequences is used to further sub-localize and quantify attached deuterium label. Integrated automation of fluidics, including sample preparation (functional deuteration), sample storage and injection (cryogenic autosampler), solid-state proteolysis, liquid chromatography, and mass spectrometry allows rapid, continuous data acquisition, typically with one sample processed every 20 minutes. With these enhancements, the present invention provides complete a high resolution, comprehensive DXMS analysis of a protein in two weeks, and can process 10 proteins simultaneously. A detailed DXMS analysis of twenty-four different *Thermotoga maritima* proteins under crystallographic study has been performed (Lesley, et al. Proc Natl Acad Sci U S A 99:11664-9. 2002) (see Examples herein). Data acquisition, deconvolution to produce exchange rate fingerprints, and detailed analysis was successfully completed for twenty-one of these proteins within a two-week period.

[0212] DXMS-derived exchange rate fingerprints. MS scans containing the numerous peptides of interest are individually isolated from the mass- intensity lists, processed to optimize signal-to-noise ratios, and then the geometric centroids of the isotopic envelopes of each peptide determined and recorded. Calculation of the difference in weight between the measured centroid of the deuterated peptide and the centroid for the same peptide without deuterium allows determination of the amount of deuterium on each peptide at the time of

MS measurement. These data manipulations are now automatically performed by specialized data reduction software (Hamuro, et al. J. Mol. Biol. 4:703- 714 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Wong, et al. Protein Science 11:73 2002). This, and other software allows efficiently processing of the tremendous amount of data obtained through automated DXMS data acquisition in a matter of hours, greatly speeding both fragmentation optimization and calculation of peptide deuterium content. The result is the determination of exchange rates for the majority of the amides in a target protein. The two-dimensional matrix formed by annotating each peptide bond amide within a protein with its hydrogen exchange rate, as measured under conditions native for the 3D-structure of the protein, constitutes its "exchange rate fingerprint".

[0213] Construction of a Spectrin Exchange-Rate Fingerprint from DXMS data. A typical example of high-density fragmentation data achieved with the methods of the present invention, and how it can be processed to produce an exchange rate fingerprint from a two-tandem- repeat construct of chicken brain α spectrin (16^{th}-17^{th} repeats) is presented in the Examples herein. The (crystal) structure of this construct had been previously determined to be a coiled coil with five very long alpha helices linked by four short loops (Grum, et al. Cell 98:523-535 1999). Spectrin was first subjected to denaturation in varying concentrations of guanidine hydrochloride (GuHCl; 0, 0.05, 0.5, 4.0 M) under quench conditions (0°C, pH 2.7,) followed by digestion with solid- state pepsin for 30 seconds. It was found that 0.5M GuHCl produced sufficient fragmentation for the initial study, and the resulting fragmentation map, consisting of 108 overlapping peptides generated. Spectrin was then on-exchanged in deuterated buffer for varying times at (10 seconds to 24 hours) at 22 °C, samples then exchange-quenched, fragmented with pepsin, fragments identified, and deuterium on each peptide, at each exchange time point quantified by the forgoing DXMS methodologies. The methods for both manual and computational deconvolution of such data into rate maps is presented in the Examples herein.

[0214] COREX-Generated Exchange Rate Fingerprints. Studies establishing the ability of COREX to calculate NMR- validated exchange rates for proteins, in this case for turkey ovomucoid third domain (2ovo) have been published (Wrabl, et al. Protein Sci 10:1032-45. 2001). A high correlation between COREX and NMR-derived exchange rates was found. To

further probe the utility of the COREX algorithm to generate a useful stability fingerprint, particularly for the purpose of assessing the relative quality of structural predictions, COREX analysis of four target proteins from the Computer Assisted Structure Prediction 2 exercise (CASP2) was performed, along with the "best" and "worst" structural predictions officially submitted for each target. A strong correlation was found for the COREX calculation of the "best" prediction and the thermodynamic signature of the target structure Conversely, the "worst" prediction showed no correlation. This result indicates that COREX analysis alone can be used to identify poor fold predictions.

[0215] Simulations of the DXMS-COREX filter. The results of the CASP4 analyses were reviewed, and two of its target proteins were selected to which at least one substantially correct structure prediction had been made, but where the remainder of the predictions demonstrated a considerable range of accuracy. One of the targets (CASP-4 target T0102 (Bacteriocin AS-48) was selected for initial study. COREX was run against each of the 87 structural models submitted for the complete structure of this target, employing an 8 residue window to produce 3,856 partially folded states. The analysis was performed on 2 nodes (sixteen 375 GHz cpu's) on BlueHorizon at the San Diego Supercomputing Center and took a total of 2 hours to run. The determined crystal structure (Bacteriocin AS-48, PDB 1E68) of target T0102 was also analyzed by COREX and its rate fingerprint calculated. The RMSD between the protection factor fingerprint of each prediction and the fingerprint derived from the crystal structure was calculated.

[0216] In Figure 1, the predictions are ranked by degree of RMSD agreement between prediction fingerprints and the actual structure's rate fingerprint, with the positions of the eight best structural predictions for this target (as determined by CASP) indicated by arrows. As can be seen, the sum of the residuals (residue-specific RMSD values) is lowest for structures that scored very well in the CASP4 contest, relative to those structures that scored poorly. The sole exception is the behavior of the second- best scoring prediction (#2) which ranked 47th in the degree of COREX- determined fingerprint similarity with that of the crystal structure. In Figure 2 the predictions in this same study are ranked by 3D structural accuracy, with the positions of the eight best "fingerprint fits" indicated by arrows. There is a very strong correlation between prediction accuracy and fingerprint fit, at least for the best

predictions. This analysis was repeated for the CASP4 protein T0125 with a window size of 10, and again, the top structure predictions had the closest agreement in exchange rate fingerprints. (Figure 3).

[0217] Use of Monte Carlo sampling to drastically reduce computing time. The computing time for COREX calculations scales exponentially with increasing protein size. This makes application of the DXMS-COREX Filter to large proteins problematic unless strategies can be developed that allow the necessary calculations to be performed within available resources. A number of approaches to mitigate this problem are readily implemented. These include improvements to the fundamental speed of the algorithm and COREX software , increased parallelization, and the use of larger computing grids. Most importantly, the methods of the present invention recognize that the goal is to perform the COREX calculations in a manner that is just sufficient for the "filter" to work, and to no higher precision.

[0218] Further study indicates that running COREX in a sparse Monte Carlo sampling mode can markedly decrease the computational time required to produce filter- targeted rate fingerprints. High-density DXMS data on a 221 amino acid construct of chicken brain alpha spectrin was acquired. On-exchange data had been collected on spectrin fragments derived from protein deuterated for 10 to 3000 seconds at room temperature (Figure 11A), and then processed to produce the construct's actual DXMS rate fingerprint (Figure 11C). The exchange rate fingerprint of the construct from its known crystal structure was calculated with COREX. The exchange rate fingerprint of the chicken spectrin construct discussed above was calculated from its known crystal structure with COREX operating in Monte Carlo mode, employing a window size of 8 and a total sampling of 8000 partially unfolded states (Figure 11C). It is readily apparent that there is a remarkably close agreement between the true (DXMS) rate fingerprint and the Monte Carlo- calculated fingerprint. Even more remarkably, this was accomplished with a run time ~24500 fold faster than would have been required for a complete, high- resolution analysis of all unfolding states. See Examples for a further description of the deconvolution methods employed in this study.

[0219] High throughput DXMS identification of amino acids on the surface of proteins to improve Rosetta predictions. The most immediately useful DXMS information for

improvement of Rosetta predictions likely comes from the identification of amide hydrogens that exchange at the maximal possible rate, indicating that they are fully solvated and on the surface of the protein. These "very fast-amides" are hydrogen-bonded to solvent water rather than the protein most of the time, and may variously be present in structured regions (short loops, in kinks in α-helices, and in edge strands of β-sheets) or disordered stretches of sequence. While it is clear that such constraints will likely be of great value in improving structural predictions these experimental constraints must be obtainable in a rapid and economical fashion if they are to be of practical use.

[0220] A recently developed DXMS-approach to rapidly localize disordered regions in proteins can also be used to localize protein surface amino acids in a high throughput, economical manner, as taught herein. As "high-throughput" DXMS is used to localize long stretches (4 or more contiguous residues) of rapidly exchanging sequence in proteins, these regions represent "disordered" regions in the protein. These disordered regions were then engineered out of the proteins to see if crystallization success was improved for use in x-ray crystallographic studies.

[0221] DXMS analysis was successfully performed on 24 *Thermatoga maritima* proteins with various crystallization and diffraction characteristics. Data acquisition was performed in a single 30 hour run, and reduction of the data to exchange rate maps was completed in two weeks, with resulting localization and prediction of several unstructured regions within the proteins. When compared with those targets of known structure, the DXMS method correctly localized small regions of disorder. DXMS analysis was then correlated with the propensity of such targets to crystallize and was further utilized to define truncations that might improve crystallization. Truncations that were defined solely on the basis of DXMS analysis demonstrated greatly improved crystallization, and were successfully used to obtain high-resolution structures for two proteins that had previously failed all crystallization attempts.

[0222] Figure 5 shows the ten second amide hydrogen/deuterium exchange map for the protein TM0449. The brief, 10 second deuteration allowed selective labeling of the most rapidly exchanging amides in the protein. The horizontal dark bars are the protein's pepsin-generated fragments that had been produced, identified, and used as exchange rate probes in the subsequent 10-second deuteration study. The number of deuterons that went on to each

peptide in 10 seconds is indicated by the number of red residues in each peptide. Two extensive segments are seen to be deuterium labeled: 1 (Phe 31-Glu 38) and 2 (Ser 88- Lys 93). Figure 5B shows the electron density of the crystal with two regions of disordered sequence, corresponding to the segments 1 and 2. Detailed electron density maps are shown in Figures 5C and 5D, in which density is not visualized between the Phe 31 to Glu 39 and Ser 88 to Ser 95 regions of the TM0449 3-D structure.

[0223] Figure 4 shows the deuteration results for all of the 21 proteins that were analyzed, whose amino acid lengths varied from 76 to 461 residues. Dark regions indicated fast exchanging amides and clear regions indicate stretches of no exchange. Regions of four or more fast exchanging amides are circled. While the circled stretches of sequence were the focus of our study, the present invention illustrates that the isolated (single to triple) rapidly exchanging amides that are peppered throughout these exchange rate maps likely represent the very rapidly exchanging amides of structured residues on the surface of the proteins. The design of this study was biased towards the detection of large stretches of rapidly exchanging sequence. Minor changes in experimental design (higher fragmentation intensity, measurement of off exchange under quench conditions) readily allow high resolution localization, and more sensitive identification of the isolated rapid exchangers. It is anticipated that 10-20 % of a typical protein's amides can be detected as being surface-localized by DXMS analysis with these simple modifications to data acquisition.

[0224] Computational economy. One way to incorporate COREX into Rosetta is to use an evolutionary-algorithm approach. In this iterative method, Rosetta , or any other computational structure predictive method is used generate a set of predictions (decoys), COREX or other amide rate calculating methods used to score them to judge their fitness, then the best-scoring decoys will be used as parents to generate a next generation of decoys. The cycle can be iterated until the COREX score converges. Many variants of this method can be employed, by altering the methods used for generating children from parents. One simple method is to raise the temperature during a Monte Carlo minimization to allow mutations to occur. That simple method has been demonstrated to efficiently minimize various fitness functions. More complex genetic algorithms could be tested as well, in which recombination between two or more is also allowed.

[0225] Alternative methods for obtaining experimental hydrogen exchange rates and calculating exchange rates from 3-D structures. While any of a number of alternative approaches to the calculation of amide hydrogen exchange rates from actual or presumed 3-D protein structures can be used in this method, a preferred one is that embodied in the COREX algorithm of Hilser and associates, and its successors, embodied in the Fyrestar software provided by Redstorm Scientific, Inc. Similarly, while the method for characterization of the fine structure of protein binding sites and for defining solvent accessible amide hydrogens employing DXMS is a preferred method for obtaining the needed hydrogen exchange rates, rates obtained in part or totally by other hydrogen-exchange methods, including those previously described in the art, as can exchange rate measurements from NMR measurements. Finally the method continues to work well even when the rate information is incomplete, and/or of low resolution.

[0226] "Multidimensional thermodynamic constraint" method. The filter approach does not make optimal use of what is the most informative aspect of DXMS rate data: precise definition of thermodynamic parameters for each residue in a protein (stability and with suitably acquired data, enthalpy) that can be put to use in predicting structure. An additional aspect of the present invention, the "multidimensional constraint" extension, employs DXMS data to apply multidimensional constraints to a protein's COREX-calculated amino acid thermodynamic environmental propensities thereby allowing DXMS data to refine protein structure prediction/determination. The ability of COREX to calculate preferred environments of (i) stability, and (ii) enthalpy has been described, and has been used to calculate and identify fold- specific stability/ enthalpic fingerprints based on primary sequence alone, as mentioned above. In the present invention, suitably obtained DXMS data on small amounts of the target protein (*e.g.* less than about 10 micrograms) are employed to further constrain estimates of residue-specific environments, refining the precision of the resulting fold-predictions. The output of this process can optionally be further evaluated with the use of the filter approach described herein.

[0227] The invention will now be described in greater detail by reference to the following non-limiting examples.

EXAMPLE 1

Rapid Refinement of Crystallographic Protein Construct Definition Employing Enhanced Hydrogen/Deuterium Exchange Mass Spectrometry (DXMS)

[0228] It is widely anticipated that access to high-resolution protein structures will be facilitated by novel high-throughput improvements to conventional crystallographic methods. Proteome-wide crystallography is one avenue being pursued by several groups, including the Joint Center for Structural Genomics (JCSG) (Lesley, et al. Proc Natl Acad Sci U S A 99:11664-9. 2002, Stevens, et al. Science 293:519-520 2001, Stevens, et al. Science 294:89-92 2001). These efforts have benefited greatly from recent technology enhancements in protein expression and crystallization. Despite these enhancements, production of stable proteins that produce suitable crystals continues to be a serious bottleneck. Many generally well-structured proteins contain disordered regions that may serve as passive linkers between structurally autonomous domains, or become ordered when they interact with binding partners that provide stabilizing atomic contacts (Wright, et al. Journal of Molecular Biology 293:321-331 1999). Regardless of their function, unstructured regions can inhibit crystallization. Unstructured regions of proteins are also particularly susceptible to contaminating cellular proteases. Removing disordered regions may improve homogeneity. The energetics and kinetics of protein crystallization may be facilitated by selective deletion of unstructured sequences (Kwong, et al. J. Biol. Chem. 274:4115-4123 1999). Even those proteins that readily crystallize can suffer from poor diffraction, and it is likely that disorder plays a significant role. Truncated constructs should result in better diffraction and, consequently, result in higher resolution data more amenable to automated map fitting procedures (Cohen, et al. Protein Science 4:1088-1099 1995, Lamzin, et al. Nature Structural Biology November 7, 2000, Supplement:978-981 2000).

[0229] In principle, information regarding protein dynamics could be used to design truncations that retain structure and maintain biological function but are otherwise depleted of disordered regions. A number of approaches ranging from stability-dependent protein expression screens to computation of stability from primary structure have been reported (Dunker, et al. Pac. Symp. Biocomput. 3:473-484 1998, Garner, et al. Genome Inform. 9:201-214 1998, Romero, et al. Pac. Symp. Biocomput. 3:437-448 1998). For structural genomics studies, many targets have unknown folds, which limits the utility of bioinformatic

predictions. NMR spectroscopy is one of the most powerful techniques to provide protein dynamics information, however, protein quantity, concentration, experimental time, and size are often limiting factors. Though limited proteolysis coupled to mass spectrometry is a preferred approach, its use is time consuming, frequently requiring that multiple proteolytic reactions be refined for optimal cleavage (Cohen, et al. Protein Science 4:1088-1099 1995). Interpretation of limited proteolysis results is confounded by the possibility that proteolysis may clip internal loops, leading to destabilization and further proteolytic degradation of what originally was a structured region. Most importantly, there is no facile method to confirm that the truncations designed have retained the stable elements of the full-length protein. These approaches are problematic in structural genomics efforts, where throughput and cost are dominating considerations (Chen, et al. Protein Science 7:2623-2630 1998).

[0230] For more than 40 years, peptide amide hydrogen-exchange techniques have been employed to study the thermodynamics of protein conformational change and the mechanisms of protein folding (Englander, et al. Methods Enzymol. 232:26-42 1994, Bai, et al. Methods Enzymol. 259:344 1995). More recently, they have proven to be increasingly powerful methods by which protein dynamics, domain structure, regional stability and function can be studied (Englander, et al. Protein Science 6:1101-9 1997 , Engen, et al. Analytical Chemistry 73:256A-265A 2001). Deuterium exchange methodologies coupled with Liquid Chromatography Mass Spectrometry (LCMS) presently provide the most effective approach to study exchange rates in proteins (Engen, et al. Analytical Chemistry 73:256A-265A 2001). Proteolytic and/or collision-induced dissociation (CID) fragmentation methods allow exchange behavior to be mapped to subregions of the protein (Engen, et al. Analytical Chemistry 73:256A-265A 2001, Hoofnagle, et al. Proceedings, National Academy of Sciences 98:956-961 2001, Resing, et al. J. Am Soc Mass Spectrom 10:685-702 1999, Mandell, et al. Anal. Chem. 70:39487-3995 1998, Mandell, et al. Proc Natl Acad Sci U S A 95:14705-10. 1998, Mandell, et al. J. Mol. Biol. 306:575-589 2001, Kim, et al. J Am Chem Soc 123:9860-6. 2001, Kim, et al. Biochemistry 40:14413-21. 2001, Zhang, et al. Protein Sci 10:2336-45. 2001, Kim, et al. Protein Sci 11:1320-9. 2002, Peterson, et al. Biochem J 362:173-81. 2002, Yan, et al. Protein Sci 11:2113-24. 2002). Building upon the pioneering work of Englander and Smith (Englander, et al. Protein Science 6:1101-9 1997, Engen, et al. Analytical Chemistry 73:256A-265A 2001, Smith, et al. J. Mass Spectrometry 32:135-146

1997), the present invention has developed and implemented a number of improvements which have significantly improved throughput, comprehensiveness, and resolution. The methods employing these enhancements high-throughput and high-resolution have been termed Deuterium Exchange-Mass Spectrometry (DXMS) (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Hamuro, et al. J. Mol. Biol. 4:703- 714 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Woods-Jr. U.S. Patent No. 5,658,739 (1997), Woods-Jr. U.S. Patent No. 6,291,189 (2001), Woods-Jr. U.S. Patent No. 6,331,400 (2001), Woods-Jr. U.S. Patent No. 6,599,707 (2003), Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003).

[0231] Peptide amide hydrogens are not permanently attached to proteins, but reversibly interchange with hydrogen present in solvent water. The chemical mechanisms of the exchange reactions are understood, and several well-defined factors can profoundly alter exchange rates (Englander, et al. Methods Enzymol. 232:26-42 1994, Englander, et al. Anal. Biochem. 147:234-244 1985, Englander, et al. Methods Enzymol. 26:406-413 1972, Englander, et al. Methods Enzymol. 49G:24-39 1978). One of these factors is the extent to which a particular exchangeable hydrogen is exposed (accessible) to water. In a completely unstructured polypeptide sequence, peptide amide hydrogens are always maximally accessible to water and exchange at their maximal rate, which is approximately (within a factor of 30) the same for all amides; their half-life of exchange is in the range of one second at 0° C and pH 7.0 (Molday, et al. Biochemistry 11:150 1972, Bai, et al. Proteins: Structure, Function, and Genetics 17:74-86 1993). Most amide hydrogens in structured peptides or proteins exchange much more slowly (up to 10^9 - fold reduction), reflecting the fact that exchange occurs only when transient unfolding fluctuations fully expose the amides to solvent water. The exception is the set of very fast exchanging amides in structured regions that have their amides fully solvent- exposed at all times, reflecting their protein-surface disposition. In effect, each amide's exchange rate in a native protein directly and precisely reports solvent accessibility to it, thereby revealing the protein's thermodynamic stability on the scale of individual amino acids. Measurement of the exchange rates of a protein's amides can therefore allow direct identification and localization of structured/ unstructured regions of the protein; unstructured regions are those where substantial contiguous stretches of primary

sequence exhibit the maximal possible exchange rates, indicative of complete and continuous solvation of the amide hydrogens in such segments (Englander, et al. Methods Enzymol. 232:26-42 1994, Bai, et al. Methods Enzymol. 259:344 1995). With its high-throughput capabilities, DXMS can rapidly localize disorder within crystallographic targets using a minimum of protein sample.

[0232] One aspect of the present invention focuses on proteins from *Thermotoga maritima* (Lesley, et al. Proc Natl Acad Sci U S A 99:11664-9. 2002). An unbiased set of *T. maritima* targets, 1376 of the 1877 predicted open reading frames, were processed through expression and purification attempts. Of these, 542 proteins were expressed in soluble form and setup for crystallization trials with 434 resulting in preliminary crystal hits. This large dataset provides the basis to select proteins for DXMS analysis based on their propensity to crystallize. To sharply focus this analysis, a subset of *T. maritima* proteins selected for their range of known crystallization behavior were investigated. The methods of the present invention use DXMS to improve crystallographic construct design under high-throughput conditions.

Protein expression and purification

[0233] Twenty-four *T. maritima* proteins were selected for analysis (see Table 1 below). These proteins, and the subsequently designed truncated constructs, were freshly prepared for this study as previously described (Lesley, et al. Proc Natl Acad Sci U S A 99:11664-9. 2002). In brief, all targets were expressed in either *E. coli* DL41 or HK100 from plasmids based on the expression vector pMH1 or pMH4. These vectors encode a 12 amino acid tag containing the first 6 amino acids of thioredoxin and 6 His residues placed at the N-terminus. Expression was induced by the addition of 0.15% arabinose for 3 hours. Bacteria were lysed by sonication, cell debris pelleted, and proteins purified from the soluble fraction by nickel chelate chromatography. Proteins were concentrated to a final volume of 0.75 μl with concentrations ranging from 15 to 50 mg/ml in 20mM TrisHCl, pH8.0 with 150 mM NaCl (Lesley, et al. Proc Natl Acad Sci U S A 99:11664-9. 2002).

Establishment of protein fragmentation probe maps

[0234] Aliquots of each of the 24 proteins were adjusted to a concentration of 10 mg/ml in Tris-Buffered Saline (5 mM Tris, 150 mM NaCl, pH 7.0; TBS), and all subsequent steps performed at 0° C on melting ice. In a 4° C cold room, five μl of each solution was further diluted with 15 μl of TBS in a microtiter plate employing multichannel pipettors for simultaneous manipulation. Thirty microliters of a stock "exchange quench" solution (0.8% formic acid, 1.6 M GuHCl) was then added to each sample (final concentration 0.5% formic acid, 1.0 M GuHCl), samples transferred to autosampler vials, and then frozen on dry ice within one minute after addition of quench solution as previously described (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Hamuro, et al. J. Mol. Biol. 4:703- 714 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Woods-Jr. U.S. Patent No. 5,658,739 (1997), Woods-Jr. U.S. Patent No. 6,291,189 (2001), Woods-Jr. U.S. Patent No. 6,331,400 (2001), Woods-Jr. U.S. Patent No. 6,599,707 (2003), Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003). Vials with frozen samples were stored at – 80 °C until transferred to the dry ice-containing sample basin of the cryogenic autosampler module of a DXMS analysis apparatus designed and operated as previously described (Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003). In brief, samples were melted at 0 °C, proteolyzed for 16 seconds by exposure to immobilized pepsin, fragments collected on a c18 HPLC column, with subsequent acetonitrile gradient elution. Column effluent was analyzed on both a Thermo Finnigan LCQ electrospray mass spectrometer and a Micromass Q-Tof mass spectrometer, as previously described (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Hamuro, et al. J. Mol. Biol. 4:703- 714 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Woods-Jr. U.S. Patent No. 6,599,707 (2003), Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003). The Sequest software program (Thermo Finnigan Inc) identified the likely sequence of the parent peptide ions and these tentative identifications were confirmed with specialized DXMS data reduction software as previously described (Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003).

On-exchange deuteration of proteins

[0235] After establishment of fragmentation maps for each protein, amide hydrogen exchange-deuterated samples of each of the 24 proteins were prepared and processed exactly as above, except that 5 μl of each protein stock solution was diluted with 15 μl of Deuterium Oxide (D_2O) containing 5 mM Tris, 150 mM NaCl, pD (read) 7.0, and incubated for ten seconds at 0 °C on melting ice before quench and further processing. Data on the deuterated sample set were acquired in a single automated 30-hour run and subsequent data reduction performed with the DXMS software. Corrections for loss of deuterium label were made as previously described. (Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003). The total time elapsed for data acquisition and analysis (both fragmentation maps and deuteration study) was two weeks. A total of 100 μg of each protein was used to complete the study. For subsequent comparative analysis of the exchange rates of amide hydrogens within truncated protein constructs versus their full-length forms, both proteins were contemporaneously on-exchanged as above, but quenched at varying times (10, 30, 100, 300, 1000, 3000, 10,000, and 30,000 seconds), and further processed as above, employing the fragmentation maps established for the full-length protein.

Protein crystallization and diffraction data acquisition

[0236] Proteins were crystallized using the vapor diffusion method with 50 nl or 250 nl protein and 50 nl or 250 nl mother liquor respective volumes as sitting drops on customized 96 well microtiter plates (Greiner). Each protein was setup using 480 standard crystallization conditions (Wizard I/II, Wizard Cryo I/II [Emerald Biostructures], Core Screen I/II, Cryo I, PEG ion, Quad Grid [Hampton Research]) at 4° and 20 °C. Images of each crystal trial were taken at least twice, typically at 7 and 28 days after setup with an Optimag Veeco Oasis 1700 imager. Each image was evaluated using a crystal detection algorithm and scored for the presence of crystals (Spraggon, et al. Acta. Cryst. D. 58:1915- 1923 2002). Images at days 7 and 28 were also evaluated manually. Diffraction data were provided by the JCSG from automated data collection at 100K on beamlines of the SSRL Structural Molecular Biology/Macromolecular Crystallography Resource, and the Advanced Light Source

beamlines 5.0.2 and 5.0.3 as described previously (Lesley, et al. Proc Natl Acad Sci U S A 99:11664-9. 2002).

DXMS defines rapidly-exchanging regions of *T. maritima* proteins

[0237] In DXMS analysis, fragmentation parameters are initially optimized, including denaturant (GuHCl) concentration, protease type(s), proteolysis duration to maximize the number of peptide fragment probes available for use with the target protein, and then the protein is examined using a broad range of on-exchange times. This approach optimizes the ability to measure the widely ranging exchange rates for most of the peptide amides in the protein (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Hamuro, et al. J. Mol. Biol. 4:703-714 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Woods-Jr. U.S. Patent No. 6,599,707 (2003), Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003). In the methods of the present invention, only disordered amides that exchanged very fast in the native protein were localized. Based on prior experience, a single set of fragmentation conditions was employed and on-exchanged samples for a single, brief (10 seconds, 0 °C) interval to selectively label only the most rapidly exchanging amides.

[0238] Generation of fragmentation maps and acquisition and analysis of deuteration data were completed in two weeks time for 24 samples. Fragmentation maps covering the entire protein sequence were obtained for sixteen proteins, nearly complete coverage for five proteins, and inadequate coverage for three proteins (see Table 1 below). Deuterium on-exchange studies were performed on the 21 proteins that had generated useful fragmentation maps (see Figure 4). Deuterium labeling was manually assigned to residue positions within the protein by first optimizing consensus in deuterium content of overlapping peptide probes, followed by further clustering of labeled amides together in the center of unresolved regions, so that a consensus map was generated. The deduced 10 second exchange maps for each of the 21 proteins, and their consensus maps are summarized in Figure 8.

[0239] The duration of labeling (10 seconds) was calculated to be sufficient to selectively deuterate primarily freely-solvated amides (Molday, et al. Biochemistry 11:150 1972, Bai, et al. Proteins: Structure, Function, and Genetics 17:74-86 1993). This was confirmed by first

fragmenting reference proteins with pepsin to yield unstructured peptides, followed by deuterium-exchange labeling of the resulting peptide mix for 10 seconds at pH 7.0 ,0 °C as above and, then, quenching and subjecting the mixture to DXMS analysis, but without repeated proteolysis. Under these conditions, all peptides were saturation- labeled with a 10 second period of on-exchange.

DXMS correctly localizes disordered regions in control proteins with known 3-D structures

[0240] Interpretation of the exchange maps of the *T. maritima* proteins was guided by the expectation of two patterns of fast exchange labeling: structurally stable, but well solvated, rapidly exchanging residues (one to three contiguous residues) versus labeling of longer stretches of sequence (four or more residues) indicative of disorder. It was presumed that three contiguous amino acids was likely the smallest number needed to complete a structurally stable turn on the surface of a protein. The percent of each protein's residues that rapidly labeled in stretches of four or more residues is indicated as "DXMS%" in Table 1.

[0241] The structure of *T. maritima* thyl protein TM0449 has been determined to 2.25 Å (Mathews, et al. Structure 11:677- 690 2003). Its exchange map demonstrated two segments ($\geq$4 residues in each) with rapid exchange, labeled A (Phe 31- Glu 38) and B (Ser 88- Lys 93), and several isolated rapidly exchanging amides in groups of 3 or less, scattered throughout the sequence (see Figure 5). Both of the rapidly-exchanging segments corresponded closely to regions of disorder in the crystal (Phe 32 - Glu 38 and Ser 89- Ser 94, Figure 5) confirming the ability of DXMS data to detect and localize such disordered regions. Interestingly, these regions also appear to be involved in the binding of the enzyme substrate and adopt a structured conformation after binding ligand (Mathews, et al. Structure 11:677-690 2003). This suggests that DXMS can also provide some localized prediction of substrate and cofactor binding sites. This raises the caution that even focused deletion of unstructured regions always carries the potential to remove regions critical to biological function. Similar comparisons were performed for other proteins with known structures (data not shown) with regions of internal disorder typically mapping to loop or extended solvent- accessible regions.

Poorly crystallizing *T. maritima* proteins contain substantial disorder

[0242] The exchange map for *T. maritima* GroES heat shock protein TM0505 demonstrated rapid exchange for three segments containing four or more contiguous rapidly-exchanging residues, which together constitute 16 % of its sequence (Figure 6). While this *T. maritima* protein had previously produced only poorly diffracting crystals, it is a close homolog of the GroES heat shock protein of *M. tuberculosis*, for which crystal structures were available as the GroES heptamer, and as a complex (GroELS) with the GroEL subunit (Ranson, et al. Cell 107:869- 879 2001, Roberts, et al. J.Bacteriol. 185:2003). When the *T. maritima* residues with rapid exchange are mapped on the *M. tuberculosis* structures, they predominantly localize to disordered residues in GroES that make contact with the GroEL binding surface.

[0243] The exchange map for the conserved hypothetical protein TM1816 (Figure 8) is dominated by several substantial regions of disorder, constituting 17.7% of its residues. This protein was a unique example where a structure was obtained from a target exhibiting substantial disorder. The poorly crystallizing proteins TM1171, TM0160, TM1706, TM1733 and TM1079 exhibit, for substantial portions of their sequence, rapidly exchanging stretches of 4 or more residues (13.9%, 12.1%, 11.5%; 6.6% and 5.7% respectively, Figure 8). TM0160, TM1171, and TM1172 had disorder primarily at the carboxy-terminus (Figure 8). These targets offer a straightforward route to domain optimization by simple deletion of the disordered carboxy-terminus. The optimization of two of these targets is described below.

Disorder-depleted constructs of *T. maritima* proteins preserve ordered structure

[0244] Truncation mutants of TM0160 and TM1171 proteins were prepared (Figures 7A and 7B), in which the carboxy- terminal disordered region(s) of both proteins were deleted. The fragmentation patterns produced by pepsin often exhibited preferences for sites near exchange-defined stretches of disorder. Several truncated constructs to each full-length protein were produced, in part guided by the location of the "preferred" pepsin cut sites, and for both TM0160 and TM1171. Deletions were designed solely on the basis of DXMS experimental data. The truncations expressed well as a soluble protein. Full-length TM0160, and its longest truncated version (D3), were on-exchanged variously for 10, 100, 1,000, and

10,000 seconds at 0 °C on ice, exchange- quenched and subjected to comparative DXMS analysis as described above. The resulting 10-second exchange maps for full-length protein and the D3 truncated version (Figure 7C) had virtually identical 10 second patterns, and detailed analysis of the longer exchange times demonstrated that D3 had a stability profile identical to that of the TM0160 full-length. Similarly, each of the four TM1171 truncated constructs expressed well as soluble protein, and had DXMS stability maps identical to that of the TM1171 full-length protein in the corresponding sequence regions (data not shown).

Deletion constructs of two *T. maritima* proteins show marked improvement in crystallization

[0245] Full-length TM0160 and the D3 truncation were submitted for crystallization trials (Table 1). A total of 480 commercially available crystallization solutions were screened at 4 °C and 20°C as described herein. From multiple protein preparations and crystallization attempts the full-length protein showed marginal crystals (inadequate for diffraction experiments) for only 3 of 2400 total attempts. In contrast using the same 480 crystallization solutions, 76 crystal hits were obtained for the truncated constructs from 1920 attempts. Crystals from the TM0160 D3 truncation mutant had better morphology than did the few crystals obtained with the full-length construct and diffracted well. Ultimately, a 1.9 Å dataset from selenomethionine-incorporated protein enabled determination of the TM0160 3-dimensional structure, which represents a novel fold (to be presented elsewhere). Similarly, the TM1171 and truncations were subjected to crystallization trials. Whereas the TM1171 full-length protein again showed very marginal crystallization propensity (5 out of 2400 attempts), each of the four TM1171 deletion constructs showed marked improvement in crystallization success with the TM1171- D4 construct ultimately resulting in a 2.1 Å dataset that was used to determine its 3-dimensional structure (to be presented elsewhere). It should be noted that well- diffracting crystals were obtained for DXMS-designed deletion constructs in both native and selenomethionine forms.

[0246] **Table 1.** Description of *T. maritima* proteins studied, as classified by crystallization history. Computational predictions (SEG%) (48) and the portion of each protein's sequence found to be present in high-exchange rate stretches of primary sequence (four or more rapidly exchanging contiguous residues; DXMS%) are given as a percentage of total residues. The primary location of the DXMS- identified rapidly-exchanging regions is

indicated. The number of unique crystallization tests is indicated along with the number of tests showing crystal hits or crystals of sufficient size to mount for diffraction screening. The percentage of total tests that led to crystals is indicated. Those targets showing less than a 1% hit rate are considered poorly crystallizing. The number of crystals screened for diffraction and the best resolution are indicated where data are available.

Table 1.

					Crystallization				Diffraction	
Target	Structure	SEG %	DXMS %	Location	Screened	Hit	Mountable	% Crystallized	Screened	Resolution (Å)
TM0665	1J6N	11.2	3.1		1920	247	244	25.6	7	2.2
TM1056	pending	17.7	0.0		175	153	9.8	45	1.8	
TM0064	1J5S	3.0	1.1		2880	114	146	9.0	31	1.8
TM1464	pending	10.1	1.3		2400	71	98	7.0	68	2.6
TM1080	1O1X	3.9	2.6		2400	119	26	6.0	32	3.5
TM0449	1KQ4	2.6	6.4		1152	42	16	5.0	16	2.3
TM0486		5.7	0.0		3552	139	29	4.7	13	3.9
TM0542	pending	9.3	2.9		2592	47	49	3.7	56	2.8
TM1733		10.0	6.6	internal	1920	33	22	2.9	29	3.4
TM1158	1O1Y	2.5	2.6		2880	32	39	2.5	62	2.2
TM0269	1J6R	2.8	4.0		4608	55	33	1.9	11	2.1
TM1764		29.3	n.d.		480	4	2	1.3	0	
TM1816	1O13	11.0	17.7	internal	2016	2	21	1.1	9	2.0
TM0505		21.1	16.3	internal	3744	12	10	0.6	8	6.3
TM0212		21.3	0.0		1152	5	1	0.5	1	8.6
TM0320		30.4	0.0		1152	4	1	0.4	0	
TM1171		10.3	13.9	C-term	4	1	0.2	0		
D1		16.4	5.9		2880	41	33	2.6	0	
D2		16.5	5.9		2880	66	9	2.6	16	2.3

D3		15.3	5.6		24	10	1.8	0				
D4	pending	14.9	5.6		1920	12	14	1.4		1		2.1
TM1706		17.9	11.5	internal	1440	3	0	0.2		0		
TM1172		11.5	3.5	C-term	3	1	0.2	0				
TM0913		7.5	2.2		4320	5	2	0.2		0		
TM1079		16.3	5.7	internal	1920	3	0	0.2		0		
TM1773		10.4	n.d.		1440	1	1	0.1		0		
TM0160		15.5	12.1	C-term	2	1	0.1	0				
D3	pending	11.6	2.5		1920	37	39	4.0		3		1.9
TM0855		9.1	n.d.		1920	0	0	0.0		0		

[0247] These studies have shown that DXMS analysis can reliably detect and localize disordered regions within an otherwise structured protein. Stability profiles were determined for 21 *T. maritima* proteins that had previously been subjected to crystallization studies (Table 1). Twelve proteins crystallized readily in >1% of the conditions tested. Four of the remaining nine poorly-crystallizing proteins had a high fraction (>10%) of their sequence in disordered regions suggesting this as a potential cause of the poor behavior. Most importantly, present instrumentation allowed determination of the DXMS-protein stability profiles at speeds matching the needs of HT Structural Genomics.

[0248] The methods of the present invention have also established that successful strategies to selectively delete disorder from protein constructs can be readily discerned from DXMS stability profiles. Furthermore, the present invention shows that DXMS can rapidly and reliably assess the fidelity of preservation of full-length structure in truncations. While several bioinformatic approaches to construct design can be used with well-characterized protein folds, DXMS-guided construct redesign offers a particular advantage in the study of proteins that have novel folds. DXMS data directly localizes disorder to specific amino acid residues in the target protein regardless of overall fold structure, allowing greatly refined truncation definition. Unlike NMR methods, which can also provide exchange data, DXMS requires only microgram amounts of soluble protein and data acquisition and analysis can be performed in a rapid timescale. In the present investigation, the total time elapsed for data acquisition and analysis (both fragmentation maps and deuteration study) was two weeks, and a total of 100 μg of each protein was used.

[0249] Finally, these results establish that DXMS stability profile- guided construct design can produce derivatives of poorly crystallizing proteins that crystallize and diffract well. In each of two attempts, the methods described herein succeeded in producing diffraction quality crystals of truncated constructs of *T. maritima* full-length proteins that had behaved poorly in several crystallization attempts, and have confirmed that these truncations preserved full-length exchange rate patterns, indicating that they had retained full-length structure with high fidelity. Taken together, these results indicate that DXMS is a valuable tool for structural genomics efforts.

Sample processing for establishment of protein fragmentation probe maps

[0250] Vials with frozen samples were stored at – 80 °C until transferred to the dry ice-containing sample basin of the cryogenic autosampler module of the DXMS apparatus. Samples were individually melted at 0 °C, then injected (45 µl) and pumped through an immobilized pepsin column (0.05% TFA, 250 ul/min, 16 seconds exposure to pepsin; 66 µl column bed volume, coupled to 20AL support from PerSeptive Biosystems at 30 mg/ ml). Pepsin- generated fragments were collected onto a C18 HPLC column, eluted by a linear acetonitrile gradient (5 to 45 % B in 30 minutes; 50 µl/min; solvent A, 0.05% TFA; solvent B, 80% acetonitrile, 20% water, 0.01% TFA), and effluent directed to the mass spectrometer with data acquisition in either MS1 profile mode or data-dependent MS2 mode. Mass spectrometric analyses used a Thermo Finnigan LCQ electrospray ion trap type mass spectrometer operated with capillary temperature at 200° C or an electrospray Micromass Q-Tof mass spectrometer, as previously described (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Hamuro, et al. J. Mol. Biol. 4:703- 714 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Woods-Jr. U.S. Patent No. 6,599,707 (2003), Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003). The Sequest software program (Thermo Finnigan Inc) identified the likely sequence of the parent peptide ions. Tentative identifications were tested with specialized DXMS data reduction software developed in collaboration with Sierra Analytics, LLC, Modesto, CA. This software searches MS1 data for scans containing each of the peptides, selects scans with optimal signal-to-noise, averages the selected scans, calculates centroids of isotopic envelopes, screens for peptide misidentification by comparing calculated and known centroids, then facilitates visual review of each averaged isotopic envelope allowing an assessment of "quality" (yield, signal/noise, resolution), and confirmation or correction of peptide identity and calculated centroid (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Hamuro, et al. J. Mol. Biol. 4:703- 714 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Woods-Jr. U.S. Patent No. 6,599,707 (2003), Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003).

On-exchange deuteration of proteins

[0251] After establishment of fragmentation maps for each protein, amide hydrogen exchange- deuterated samples of each of the 24 proteins were prepared and processed exactly as above, except that 5 μL of each protein stock solution was diluted with 15 μL of Deuterium Oxide (D_2O), containing 5 mM Tris, 150 mM NaCl, pD (read) 7.0, and incubated for ten seconds at 0° C on melting ice before quench and further processing. Data on the deuterated sample set were acquired in a single automated 30- hour run, and subsequent data reduction performed on the DXMS software. Corrections for loss of deuterium-label by individual fragments during DXMS analysis (after "quench") were made through measurement of loss of deuterium from reference protein samples that had been equilibrium-exchange- deuterated under denaturing conditions as previously described (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Hamuro, et al. J. Mol. Biol. 4:703- 714 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Woods-Jr. U.S. Patent No. 6,599,707 (2003), Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003). The total time elapsed for data acquisition and analysis (both fragmentation maps and deuteration study) was two weeks, and a total of 100 ug of each protein was used to complete the study. The personnel performing the data acquisition and reduction part of the study were unaware of the identity or crystallization histories of the proteins while data were being acquired and processed. For subsequent comparative analysis of the exchange rates of amide hydrogens within truncated protein constructs vs. their full-length forms, both proteins were contemporaneously on-exchanged as above, but quenched at varying times (10, 30, 100, 300, 1000, 3000, 10,000, and 30,000 seconds), and further processed as above, employing the fragmentation maps established for the full-length protein.

Equipment configuration

[0252] The equipment configuration consisted of electrically- actuated high pressure switching valves (Rheodyne), connected to two position actuators from Tar Designs Inc., Pittsburgh, as described previously (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Hamuro, et al. J. Mol. Biol. 4:703- 714 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Woods-Jr. U.S. Patent No.

6,599,707 (2003), Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003) . A highly modified Spectraphysics AS3000 autosampler, partially under external PC control, employed a robotic arm to lift the desired frozen sample from the sample well, then automatically and rapidly melted and injected the sample under precise temperature control (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Hamuro, et al. J. Mol. Biol. 4:703- 714 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003). The autosampler basin was further thermally insulated and all but 20 vial positions were filled with powdered dry ice sufficient to keep samples colder than -45° C for 18 hours. Four HPLC pumps (Shimadzu LC-10AD) were operated by a Shimadzu SCL-10A pump controller. One produced forward flow over the pepsin column, another backflushed the protease column after sample digestion (0.05 % aqueous TFA), and two delivered solvents to a downstream HPLC column for gradient elution (A: 0.05 % aqueous TFA; B; 80% acetonitrile, 20% water, 0.01% TFA; 1 x 50 mm C18 Vydac # 218MS5105, pH 2.3). Valves, tubing, columns and autosampler were contained within a refrigerator at 2.8° C, with pepsin and HPLC columns immersed in melting ice. The timing and sequence of operation of the DXMS apparatus fluidics were controlled by a personal computer running an in-house written LabView-based program, interfaced to solid-state relays (digital input/output boards, National Instruments), controlling pumps, valve actuators, and MS data acquisition.

EXAMPLE 2

Stability of a Two Repeat Fragment of Chicken Brain α-Spectrin Probed at High Resolution by Enhanced Hydrogen/Deuterium Exchange Mass Spectrometry (DXMS): Implications for the Molecular Mechanisms of Spectrin Elasticity

[0253] Spectrin is a cytoskeletal protein involved in maintaining structural support and membrane elasticity. It includes an α-monomer of 21 tandem repeats, with each repeat composed of three well-formed, long antiparallel α-helices connected by short turns or loops, forming a "z"-shaped three-helix bundle (Grum, et al. Cell 98:523-35. 1999). It functions, in part, as an elastic molecule, demonstrating a distinctive "sawtoothed" compliance behavior, where tension remains within a relatively narrow range despite considerable lengthening. To

better understand the molecular basis of this behavior, the structural stability of α-spectrin is determined herein, at near-individual amino acid scale, with enhanced methods of peptide amide hydrogen- deuterium exchange- mass spectrometry. The behavior of a two repeat construct (R1617) of chicken brain α spectrin (16^{th}-17^{th} repeats) for which the three dimensional structure has been determined crystallographically was determined.

[0254] The construct was incubated in D2O- containing buffer for varying times, to allow "on-exchange", with solvent- accessibility- dependent incorporation of deuterium into peptide amides, and then exchange- "quenched", to effectively lock exchanged deuterium in place. The deuterium-labeled protein was then enzymatically fragmented into a large number of sequence-overlapping peptides, and further processed by LCMS to quantify deuterium exchanged onto each peptide. This data was then computationally processed into peptide amide- specific exchange rates employing novel algorithms and software described herein. The result was the generation of an amide hydrogen exchange- rate profile from which the relative thermodynamic stability or "energetic landscape" of the molecule could be assessed at the individual residue level.

[0255] Remarkably, each of the six long helices in the construct was not a uniformly stable structure, but demonstrated gradients in hydrogen exchange rates, with amides in the middle 1/4 to 1/3 portions of each helix having slow exchange rates, progressively increasing to more than 1000 times faster rates towards the ends of the helices. Additionally, the COREX algorithm was used to computationally estimate the exchange rates for the repeat from its crystal structure, and found these results to be in close agreement with the experimentally determined exchange rate profile, confirming the presence of pronounced α-helix stability gradients. Comparable helix stability gradients were not present in five other proteins.

[0256] These findings support and extend previous models of α-spectrin behavior that propose conformational rearrangements involving stretch-induced helix-loop transitions, with migration of the short loop regions back and forth along the helices. These results suggest that if this "loop-migration" model is operative in α-spectrin, then the loops will likely migrate into progressively more stable regions of the α-helices as α-spectrin is stretched,

storing mechanical energy. This energy can be recovered when the molecule relaxes, and the loop migrates back into less stable regions of the helices.

[0257] The cytoskeleton of blood cells includes many components necessary for maintaining membrane structural integrity and allowing the cells to withstand the large stresses of traversing the circulatory system. It includes tetramers of the elastic protein α-spectrin, which consists of an α-monomer of 21 tandem repeats and a beta-monomer of 16 repeats. X-ray crystal structures of constructs composed of two such tandem repeats of the α-subunit reveal that each is composed of three well-formed, long antiparallel α-helices connected by short turns or loops, forming a "z"-shaped three-helix bundle (Grum, et al. Cell 98:523-35. 1999), with the tandem repeats connected by a short α-helical "linker" region.

[0258] While v plays a critical role in the reversible deformation of the membrane, the molecular basis of this behavior, particularly its dynamic aspects, are unclear. Investigation of cloned repeats using chemical and thermal denaturation as well as atomic force microscopy (AFM) have yielded important advances in the understanding of the physical and biomechanical properties of unfolding and refolding of repeating units as well as the function of α-spectrin as a whole. Force-extension curves from AFM studies of have demonstrated that v is a highly non-linear spring with substantial relatively small peak unfolding forces (20-50pN) per repeating unit. Other studies have also indicated that these repeats unfold independently and/or in tandem (Law, et al. Biophys J 84:533-44. 2003, Rief, et al. J Mol Biol 286:553-61. 1999) with the presence of intermediates (Altmann, et al. Structure (Camb) 10:1085-96. 2002). In AFM experiments α-spectrin demonstrates a distinctive "sawtoothed" compliance behavior, where tension rises only gradually, and remains within a relatively narrow range despite considerable lengthening of the molecule.

[0259] Several models, based on crystallization and/ or atomic force microscopy studies, have been proposed to account for α-spectrin's elasticity, including tension-induced bending of the linker regions, tension- induced unwrapping or melting of the ends of α-helices into elongated loops; and catastrophic unfolding of triple helical bundles, in which the sawtoothed compliance observed is attributed to multiple tandem bundles sequentially popping open with increased tension.

[0260] A fourth mechanism has been proposed in which there is a tension-induced end-to-end lengthening of the triple helical bundles, resulting from stretch-induced migration of the short loop regions along the α-helices, accomplished by relatively little change in the total amount of helix in each bundle (Grum, et al. Cell 98:523-35. 1999). Observations presented herein support and extend this model.

[0261] Evaluation of these, and other proposed mechanisms for α-spectrin elasticity would be facilitated by a detailed characterization of the thermodynamic stability or "energetic landscape" of the α-spectrin molecule. To this end, its structural stability was probed at the individual amino acid scale employing enhanced methods of peptide amide hydrogen-deuterium exchange lc- mass spectrometry, termed DXMS. Peptide amide hydrogens are not permanently attached to a protein, but continuously and reversibly interchange with hydrogen present in water. The chemical mechanisms of the exchange reactions are understood, and several well-defined factors can profoundly alter exchange rates (Englander, et al. Methods Enzymol. 232:26-42 1994, Englander, et al. Anal. Biochem. 147:234-244 1985, Englander, et al. Methods Enzymol. 26:406-413 1972, Englander, et al. Methods Enzymol. 49G:24-39 1978). One of the factors that determines the rate of exchange is the extent to which a particular exchangeable hydrogen is exposed (accessible) to water. The exchange reaction proceeds efficiently only when a particular peptide amide hydrogen is fully exposed to solvent. Peptide amide hydrogens that are freely accessible to water exchange at their maximal possible rate, with an average half-life of exchange of approximately one second at 0 °C and pH 7.0. (Molday, et al. Biochemistry 11:150 1972, Bai, et al. Proteins: Structure, Function, and Genetics 17:74-86 1993). The precise rate of exchange of a particular fully-solvated amide can vary more than thirty -fold from this average rate, depending upon the identity of the two amino acids flanking the amide bond. Exact exchange rates expected for fully solvent-exposed amide hydrogens can be reliably calculated from knowledge of the temperature, pH and primary amino acid sequence involved (Molday, et al. Biochemistry 11:150 1972, Bai, et al. Proteins: Structure, Function, and Genetics 17:74-86 1993).

[0262] In a structured protein, most peptide amide hydrogens exchange slower (up to 10^9-fold slower) than the maximal, fully solvated exchange rate, as they are not efficiently

exposed to solvent water. Protein structure is not static, but best considered as an ensemble of transiently unfolded states: the native state ensemble. Amide hydrogen exchange occurs only when a particular transient unfolding event fully exposes an amide to solvent. The ratio of exchange rates for a particular amide hydrogen, in the folded *vs* random coil states is referred to as the exchange protection factor, and directly reflects the free energy change in the atomic environment of that particular hydrogen between unstructured and structured states of the protein. In this sense, amide hydrogens can be treated as atomic-scale sensors of highly localized free energy change throughout a protein and the magnitude of free energy change reported from each of a protein's amides in a folded vs. unfolded state is precisely equal to - RT ln (protection factor) (Bai, et al. Methods Enzymol. 259:344 1995). In effect, each peptide amide's exchange rate in a folded protein (when measured) directly and precisely reports the protein's thermodynamic stability at the individual amino acid scale (Englander, et al. Methods Enzymol. 232:26-42 1994, Bai, et al. Methods Enzymol. 259:344 1995).

[0263] Deuterium exchange methodologies coupled with Liquid Chromatography Mass Spectrometry (LCMS), presently provide the most effective approach to perform hydrogen exchange studies of proteins larger than 30 kDa in size (Engen, et al. Analytical Chemistry 73:256A-265A 2001) (Engen, et al. Analytical Chemistry 73:256A-265A 2001, Hoofnagle, et al. Proceedings, National Academy of Sciences 98:956-961 2001, Resing, et al. J. Am Soc Mass Spectrom 10:685-702 1999, Mandell, et al. Anal. Chem. 70:39487-3995 1998, Mandell, et al. Proc Natl Acad Sci U S A 95:14705-10. 1998, Mandell, et al. J. Mol. Biol. 306:575-589 2001, Kim, et al. J Am Chem Soc 123:9860-6. 2001, Kim, et al. Biochemistry 40:14413-21. 2001, Zhang, et al. Protein Sci 10:2336-45. 2001, Kim, et al. Protein Sci 11:1320-9. 2002, Peterson, et al. Biochem J 362:173-81. 2002, Yan, et al. Protein Sci 11:2113-24. 2002). Building upon the pioneering work Walter Englander and David Smith (Englander, et al. Protein Science 6:1101-9 1997, Engen, et al. Analytical Chemistry 73:256A-265A 2001, Smith, et al. J. Mass Spectrometry 32:135-146 1997), a number of improvements to their methodologies and experimental equipment that have significantly improved throughput, comprehensiveness, and resolution have been developed and implemented, collectively referred to as enhanced Deuterium Exchange-Mass Spectrometry (DXMS).

[0264] DXMS can be used to obtain sufficient information on the exchange behavior of a two repeat construct (R1617) of chicken brain α-spectrin (16^{th}-17^{th} repeats) to allow construction of a peptide amide hydrogen exchange rate map at near single-amide resolution, from which the thermodynamic stability or "energetic landscape" of the molecule could be assessed at the individual residue level. Results demonstrate that the long α-helices within the tandem repeats are not uniformly stable structures, have marked gradients in stability. If the "loop-migration" model is operative in α-spectrin, then these gradients provide the mechanism by which mechanical energy is stored in the stretched α-spectrin molecule.

Increased production of overlapping peptides of α-Spectrin construct R1617

[0265] The ability to localize and quantify detailed hydrogen exchange behavior with DXMS is largely determined by the degree to which a densely overlapping set of peptides can be proteolytically generated from the deuterated protein prior to LCMS. Prior to deuterium on-exchange analysis, digestion of exchange- quenched, undeuterated R1617 was performed on samples made to 0, 0.5, 1.0, 2.0, and 4 M GuHCl, with the duration of proteolysis with solid- state pepsin being systematically varied, to determine optimum conditions for maximally overlapping fragmentation. At 0.5 M GuHCl and 250 ul/min flow rate over the pepsin column (66 μL bed volume), 114 high quality fragments were produced. A second, higher resolution fragmentation map was also obtained by employing these conditions, but with the addition of a *Aspergillis satoi* Fungal Protease XIII (FP XIII) column (66 μL bed volume) after the pepsin column, resulting in the generation of an additional 86 peptides. A comparison of the fragments generated by pepsin and pepsin plus FPXIII is shown in Figure 9. A total of 200 fragments were obtained with the combination of the pepsin and fungal protease columns. Such extensive fragmentation and redundancy in the overlapping of peptides was essential to successful calculations of reliable exchange rates for each residue in the spectrin construct.

[0266] Once the optimal quench-compatible fragmentation conditions were established, the R1617 construct was incubated in 150 mM NaCl, 5 mM tris, pD (read) 7.0 containing 75% mole-fraction deuterated water at 22 degrees C for times varying from 3 seconds to $3.4x10^5$ seconds, and then aliquots exchange- quenched by making them to 0.5% formic acid,

0.5M GuHCl at 0 degrees C, followed by immediate cooling to and storage at – 80 degrees C. Quenched, deuterated samples were then enzymatically fragmented, and subjected to LCMS under continued quench conditions as described herein. The deuterium content of each of the 200 peptides that had been generated from each sample was then calculated from the LCMS data, for all on- exchange times, employing specialized data reduction software and corrections for back- exchange (loss of deuterium from peptides after institution of "quench") as previously described.

Construction of a low-resolution exchange rate map for spectrin R1617

[0267] Plots of deuterium accumulation for each peptide *vs* on- exchange time were constructed from data obtained by analysis of 114 pepsin-only generated peptides, as shown in Figure 10 for three representative peptides. The time axis was arbitrarily divided into three regions, (fast, medium, and slow-exchanging; Figure 10) and the number of amides on each peptide that on-exchanged deuterium in the fast, medium and slow rate classes scored. The latter class was grouped with the *very slow* class unmeasured in the limited on- exchange times ($<10^5$ sec) used in this experiment; Figure 10, *italics*). A map of rate-class *vs* construct sequence (Figure 11B) was then constructed from this information, employing a strategy in which the (generally smaller) peptides containing one rate class were first placed in amino acid sequence register, followed by placement of peptides with two, and then three, rate classes, in a manner that required that placements of the three rate classes of amides in each peptide conform with the preceding placements. The resulting "α-Spectrin Consensus Rate Map" is indicated by the arrow in Figure 11B.

[0268] This map demonstrated features that might reasonably be anticipated from the protein's structure: the short loops between the long α-helices were uniformly fast-exchanging (Figure 11B, short horizontal bars below the consensus map) while substantial regions within the α- helices were much more slowly exchanging (the locations of the long α-helices are shown as light-blue bars below Figure 11C). The map further indicated that substantial gradients in exchange rates were present within each α-helix, with slow-exchanging regions of helix gradually transitioning to fast exchanging-regions. Remarkably, these gradients in rates were not restricted to the ends of the helices, but occurred across most of their length. For example, helix B'' was fast-exchanging for its N-terminal third, medium exchanging for its middle third, and slowly exchanging for its C-terminal third.

Construction of a high-resolution exchange rate map for spectrin R1617

[0269] To study these gradients in spectrin α-helix exchange rates at higher resolution, a computational method was developed for the deconvolution of aggregate, time-dependent peptide deuteration data, to specific exchange rates for each amide hydrogen within the native protein. This method, termed "**H**igh **R**esolution, residue-specific determination of amide hydrogen exchange rates from **DXMS** data" (HR-DXMS), employs an algorithm centered on use of a two-phase numerical technique, linear programming (LP) for an initial rate estimation followed by a nonlinear least squares fit (NLS). Essential to success of the method is the derivation and incorporation of residue-specific corrections for deuterium loss during "back-exchange" in contrast to the use of "peptide-average" loss corrections usually employed in hydrogen exchange data analysis. The method can also make use of additional hydrogen exchange data, obtained by systematically varying the duration of the usually deleterious "back- exchange", to allow resolution of individual amide rates within protein regions not sufficiently resolved by enzymatic fragmentation alone. A detailed description of the method, the validation studies that have been performed, and examples of the use of its implementing software are presented below.

[0270] Figure 11C shows the results of application of HR-DXMS to the data from 200 deuterated spectrin 1617 construct fragments, obtained by the combined action of pepsin plus

FP XIII. Results are expressed as the $DG_{exchange}$ (the difference in Gibbs free energy of exchange) between the folded and unfolded form of the protein, according to equation (7)

$$DG_{exchange,i} = -RTln(k_{ex,i}/k_{int,i}) \quad (7)$$

where $k_{ex,i}$ and $k_{int,i}$ are the experimental and intrinsic (random coil) exchange rates at amide i as determined from the intrinsic rates of random coil model peptides (Molday, et al. Biochemistry 11:150-8. 1972, Bai, et al. Proteins 17:75-86. 1993).

[0271] To facilitate comparison with the low resolution consensus rate map in Figures 11B and 11C, Figure 10 is divided by two horizontal dashed lines that are placed at $DG_{exchange}$ values corresponding to the arbitrary rate divisions imposed in the generation of the approximate rate map. There is considerable agreement between the results of the two methods and the computational approach resulted in a more finely detailed and less subjective description of the exchange rate distribution within the α-helices, clearly demonstrating the extensive exchange rate gradients that traverse the helices. The A' and A'' helices have gradients with a stable central region that decreases in stability towards each end, while the B' and B" helices demonstrate more monotonic gradients with stable C-termini that gradually become less stable at the N-terminus. The tandem-repeat linker region, which, is seen to be an α-helix in the crystal structure, has a distinctly lower stability than the amides of the helices that immediately adjoin it, helix C' and A".

Calculation of the hydrogen exchange rate map of α-spectrin R1617 from its crystallographically determined structure

[0272] The experimentally determined exchange rate map for α-spectrin 1617 with purely computational estimates of hydrogen exchange rates that can be obtained with use of the COREX algorithm. COREX (implemented in the Fyrestar software of Redstorm Scientific, Houston TX), is a computational tool that utilizes the high-resolution structure of a protein as a template to generate a large ensemble of incrementally different conformational states. COREX represents proteins as ensembles of conformations rather than as discrete structures, and has been shown to predict amide hydrogen exchange rates with remarkable accuracy and precision when tested against available NMR-derived experimental data,

suggesting that the calculated ensemble captures the general features of the actual ensemble, and thus provides a realistic physical description of proteins(Hilser, et al. Proteins 27:171-83 1997, Hilser, et al. J Mol Biol 262:756-72 1996, Hilser, et al. Proc Natl Acad Sci U S A 95:9903-8 1998, Hilser Methods Mol Biol 168:93-116 2001). COREX was run in a sparse Monte Carlo mode against the structural coordinates of the α-spectrin R1617 construct and hydrogen exchange rate protection factors were calculated as described herein. Figure 12B overlays the COREX-calculated and experimentally-determined protection factor maps deduced for α-spectrin R1617 by DXMS analysis. There is significant agreement in the overall pattern of stability between the experimental and computationally derived protection factor profiles, both confirming the α-spectrin α-helix stability gradients and cross-validating the ability of HR-DXMS to derive protection factor maps that substantially match those that can be obtained computationally by COREX analysis of known three dimensional structures.

The substantial gradients in helix stability are uniquely present in α-spectrin R1617

[0273] While it was anticipated that a few amides near the end of each helix (turn residues) might exchange faster than the bulk of the helix, the gradients in exchange rates, and corresponding gradients in helical stability, extended across most of the length of each the six α-helices in the two-repeat α-spectrin construct. To evaluate the significance of these stability gradients, the hydrogen exchange rates of α-helices within five other proteins was surveyed for which experimental measurements were available: horse cytochrome c, BPTI, SNASE, HEWL, and equine lysozyme (Milne, et al. Protein Sci 7:739-45. 1998, Radford, et al. Proteins 14:237-48 1992, Loh, et al. Biochemistry 32:11022-8 1993, Kim, et al. Biochemistry 32:9609-13 1993). Typically, the exchange rates of the first 4 amides of the N-termini of helices in these proteins could not be determined by NMR, indicating that exchange occurred too rapidly to be experimentally determined at the shortest time point experimentally accessible (generally 2 minutes). This is expected since the first 4 amino acids at the N-terminus of a typical α-helix (known as the N-cap residues) do not usually have robust cis-hydrogen bond acceptors. Amino acids interior to the N-cap residues in these proteins had typical free energies of hydration between 6-8 kcal/mole, values that were found only in the linker and most stable central portions of the helices in α-spectrin R1617. The N-

terminus of the B'' helix in R17 showed values well below 6 kcal/mole fully 15 residues into the helix.

[0274] These values indicate that substantial portions of the α-spectrin R1617 helices are much less stable under solution conditions than the comparable regions of the α-helices in the five comparison proteins. Although the conformational helix-loop transitions in the crystal structure are located in the BC loop of R17 it cannot be excluded that there may be structural differences in crystalline versus solution phases of R1617. Nevertheless, the asymmetric pattern of stability at the ends of the helices provides further support for potential conformational rearrangements in these regions.

[0275] In this study, high-resolution protein stability profiles were derived for the prototypic α-spectrin two tandem-repeat R1617 by a novel experimental approach (HR-DXMS) and by use of the well-validated COREX algorithm operating on the 3-D structure of α-spectrin R1617. The two independently-derived profiles were highly concordant and demonstrated marked, unanticipated gradients in the stability of the several long α-helices in the construct. The discovery of these gradients has important implications for proposed mechanisms of α-spectrin elastic behavior.

Spectrin elongation is mediated by tension- induced catastrophic unfolding

[0276] Atomic force microscopy measurements of α-spectrin constructs reveal that repeats abruptly unfold at forces of 20 – 50 pN(Law, et al. Biophys J 84:533-44. 2003). When several repeats are in tandem, the tension- length relationship exhibits a distinctive "sawtoothed" behavior in which tension gradually rises with increasing length until an abrupt drop in tension occurs, returning almost to baseline, followed immediately by repeated tension rise and collapse with continued elongation. The result is that the α-spectrin molecule can be elongated up to multiples of its resting length, with tension constrained to a constant, relatively narrow, range. The abrupt drops in tension have been attributed to catastrophic unfolding of individual repeats, and there is considerable evidence to support this model.

[0277] However, the mechanism responsible for the short- range rise in tension with each "sawtooth" is less clear. The force required to "snap open" the repeats is well above that typically found to be exerted on the α-spectrin molecule in simulations of membrane deformation, which are more in the range of 5-10 pN. Taken together, these observations indicate that the mechanisms that account for the rise in tension with each "sawtooth" are central to understanding α-spectrin's elasticity. Studies have indicated that repeats may undergo more subtle conformational changes that mediate elasticity in this 5-10 pN regime before catastrophic unfolding of the same or tandem repeats occurs with higher tension(Rief, et al. J Mol Biol 286:553-61. 1999, Altmann, et al. Structure (Camb) 10:1085-96. 2002).

α-Spectrin elastic behavior requires efficient storage of mechanical energy

[0278] Models for α-spectrin elastic behavior should explain how mechanical energy is stored by tension- induced conformational change so as to allow efficient, low hysteresis recoil when tension is released. Models have been proposed in which tension induces gradual unwinding or melting of the ends of α-helical regions into elongated, relatively disordered loops(Altmann, et al. Structure (Camb) 10:1085-96. 2002, Paci, et al. Proc Natl Acad Sci U S A 97:6521-6. 2000). In both catastrophic unfolding, and helix- melting models it is unclear how mechanical energy could be stored without undue hysteresis: the forces that mediate the non-covalent binding interactions within the structures of proteins operate over short distances and once the distances are exceeded, the forces in large part disappear. Once tension is released, entropic forces may allow reassembly of the unwrapped helical regions, but with resulting large losses in the mechanical energy that unwrapped them.

α-Spectrin elasticity may be mediated by energy-storing loop migration

[0279] Crystal structures have been determined for two-repeat constructs consisting of the same sequence as α-spectrin R1617, but with small variations in the particular N- and C-terminal residues chosen to begin and end the construct: ie having differing "phases". These phase-differing constructs demonstrated discrete differences in their structures when crystallized (Grum, et al. Cell 98:523-35. 1999). These differences indicated that α-spectrin preferred to reduce its end-to-end distance (in the course of crystal-packing) by reorientation of the repeats by helix-loop-helix transitions that shifted the sequence- position of the short

loops connecting helices without overall change in the amount of sequence in loop or helix: tension- induced loop migration (Grum, et al. Cell 98:523-35. 1999). This model is further supported by studies of the crystallized 16th repeat of Drosophila α-spectrin which showed a conformational rearrangement of a loop region into a helix (Yan, et al. Science 262:2027-30 1993).

[0280] This model is particularly appealing, as it proposes that the tension- induced conformational changes do not substantially alter total amount of short- range binding interactions within the repeats: the fraction of sequence in loops vs helix remains constant, providing the opportunity for low-hysteresis conformational change. These results suggest that if this "loop-migration" model is operative in α-spectrin, then the loops will migrate into progressively more stable regions of the α-helices as α-spectrin is stretched, with reformation of less stable helix behind them, storing mechanical energy. This energy can be recovered when the molecule relaxes, and the loop migrates back into less stable regions of the helices, allowing reformation of the stable helical regions.

α-Spectrin elasticity may be mediated through linker-region flexibility.

[0281] The linker between the 16th and 17th repeating units of R1617, and the linker-abutting sequence in each unit (helix C' and A'') are present in the crystal structure as a single, very long uninterrupted α-helix. The phase-difference crystallization studies noted above also demonstrated that crystal packing could induce a slight bending of the linker region. This observation suggested that in solution, the linker region might be significantly more flexible than other α-helical regions of the molecule. The data support this inference, as it was found that a significant decrease in the $\Delta G_{exchange}$ (higher exchange rate) of 2-4kcal/mol in the linker region when compared to the more stable helical regions flanking the linker, supporting the idea that the linker may be intrinsically less stable in solution, despite appearing as an α-helix in the crystal structure. It is unclear though whether this decrease in $\Delta G_{exchange}$ is due to enthalpic or entropic contributions at the linker region. Presumably the flanking helices, C' and A'', are enthalpically more stable due to the hydrophobic packing of the triple helical bundle and are less vulnerable to the dynamic processes governing helical-coil transitions in single helices. The lower $\Delta G_{exchange}$ in the linker region is not due to the

varying degree of amide solvent exposure if the linker exists in solution as an α-helix, amide hydrogens are efficiently hydrogen bonded with the carbonyls in the preceding turn. Only if the amide hydrogen bond is broken and the hydrogen is exposed to the solvent can exchange occur.

[0282] These exchanges in the linker region occur may occur via local or global unfolding. If the $\Delta G_{exchange}$ of the linker region were similar to the $\Delta G_{exchange}$ of the amides in the flanking helices, then global unfolding would be the likely mechanism. However, the decrease of 2-4kcal/mol in the linker region indicates that predominately local unfolding is occurring there. Thus the helical linkers are unusually dynamic in solution, and likely exhibit a significantly increased flexibility relative to other parts of the molecule. Although the presence of an α-helix joining adjacent repeats may appear to be a "stiff" linkage, these results support the inference drawn from comparative analysis of differently "phased" structures, that under solution conditions, these linkers are dynamic structures that can provide significant configurational entropy for α-spectrin chains.

High resolution DXMS

[0283] In the course of this study, novel methods ("HR-DXMS") were developed by which high quality amide hydrogen/ deuterium mass spectrometry data can be computationally resolved into near single-amide resolution hydrogen exchange rate profiles for an entire protein construct. The method was validated by demonstrations of its ability to accurately deconvolute realistically-simulated raw DXMS fragmentation data, employing fragmentation densities, ranges of on- exchange times and data precision routinely obtained with DXMS analysis. It was further validated by demonstration of its ability to derive an exchange rate profile for α-spectrin R1617 that substantially matched that produced by analysis of the crystal structure of α-spectrin R1617 with the COREX algorithm. With this capability, HR-DXMS now rivals the resolution of NMR- based methods for amide hydrogen exchange rate measurement, with the substantial advantages of being able to measure even the fastest exchanging amides, and to do this with substantially larger proteins, and less material than is required for NMR approaches. The principal requirement for application of this method is that high quality, high fragment-density exchange data be obtained for the

study protein. This is now readily available through application of the enhanced data acquisition methods ("DXMS") recently reported (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Hamuro, et al. J. Mol. Biol. 4:703- 714 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Woods-Jr. U.S. Patent No. 6,599,707 (2003), Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003).

[0284] There is a further, unique advantage of the HR-DXMS approach: The functional deuteration step of the analysis can be performed under conditions where the dynamic properties of the study protein can be directly manipulated, with experimental design limited only by the ingenuity of the investigator. For example, with the present α-spectrin study as a foundation, one can test the "loop migration" hypothesis, and other models for α-spectrin elasticity by performing comparative HR-DXMS studies with the deuteration step performed while the α-spectrin molecule is being progressively elongated, for example by shear stress fields generated by fluid flow or stirring. After induction of quench, exactly the same method of analysis that was employed in the present study would allow rigorous assessment of the several proposed mechanisms for α-spectrin elasticity.

[0285] The COREX algorithm was developed with the goal of representing the ensemble thermodynamic behavior of proteins in a computationally accessible manner. It scales well when implemented in a (massively) parallel manner, as opposed to typical molecular dynamics calculations. The amide hydrogen exchange-rate calculating ability of COREX was originally developed to allow validation of the stability profiles it generated by comparison with NMR- derived exchange rate measurements. The rate-calculating ability of COREX will play an important role in the manner in which HR-DXMS – derived protein stability profiles and exchange rate maps are interpreted and exploited. The close agreement between HR-DXMS and COREX- derived exchange rate profiles for α-spectrin R1617 has heightened this expectation. There are, however, minor portions of the COREX-derived profile that deviate from the experimental profile: the linker region and a portion of the C-terminal region of the molecule are shown to be more stable by COREX analysis than by HR-DXMS analysis. These differences may result from an inadequate sampling of states in the Monte Carlo mode employed.

[0286] Establishment of protein fragmentation maps. Thirty microliters of stock "exchange quench" solutions (0.8% formic acid, 0M/.8M/1.6M/3.2M/6.4M GuHCl) was added to 20μL of sample (final concentration 0.5% formic acid, 0M/.05/1.0/2M/4M M GuHCl) containing 10-15ug of protein in TBS, transferred to autosampler vials, and then frozen on dry ice within one minute after addition of quench solution. Vials with frozen samples were stored at –80 deg C until transferred to the dry ice-containing sample basin of the cryogenic autosampler module of the DXMS apparatus. Samples were individually melted at 0 deg C, then injected (45 ul) and pumped through protease columns (0.05% TFA, 250ul/min, 16 seconds exposure to protease). Proteolysis used immobilized pepsin (66 μl column bed volume, coupled to 20AL support from PerSeptive Biosystems at 30 mg/ ml) or similarly immobilized Aspergillus satoi Fungal Protease XIII (20mg/ml, 66μL bed volume column) . Protease- generated fragments were collected onto a C18 HPLC column, eluted by a linear acetonitrile gradient (5 to 45 % B in 30 minutes; 50 μl/min; solvent A, 0.05% TFA; solvent B, 80% acetonitrile, 20% water, 0.01% TFA), and effluent directed to the mass spectrometer with data acquisition in either MS1 profile mode or data-dependent MS2 mode. Mass spectrometric analyses used a Thermo Finnigan LCQ electrospray ion trap type mass spectrometer operated with capillary temperature at 200 °C or an electrospray Micromass Q-Tof mass spectrometer, as previously described (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Hamuro, et al. J. Mol. Biol. 4:703- 714 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Woods-Jr. U.S. Patent No. 6,599,707 (2003), Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003). The Sequest software program (Thermo Finnigan Inc) was used to identify the likely sequence of the parent peptide ions. Tentative identifications were tested with specialized DXMS data reduction software developed in collaboration with Sierra Analytics, LLC, Modesto, CA. This software searches MS1 data for scans containing each of the peptides, selects scans with optimal signal-to-noise, averages the selected scans, calculates centroids of isotopic envelopes, screens for peptide misidentification by comparing calculated and known centroids, then facilitates visual review of each averaged isotopic envelope allowing an assessment of "quality" (yield, signal/noise, resolution), and confirmation or correction of peptide identity and calculated centroid (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Hamuro, et al. J. Mol. Biol. 4:703- 714

2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Woods-Jr. U.S. Patent No. 6,599,707 (2003), Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003).

[0287] On-exchange deuteration of proteins. After establishment of fragmentation maps amide hydrogen exchange-deuterated samples of R1617 were prepared and processed exactly as above, except that 5 μL of each protein stock solution was diluted with 15 μL of Deuterium Oxide (D2O), containing 5mM Tris, 150mM NaCl, pD (read) 7.0, and incubated at 22 degrees C. for 3,10,30,100,300,10^3,$3x10^3$,10^4,$2.5x10^5$, $3.4x10^5$ seconds, at which time samples were supplemented with 30 μL of a quench solution (0.8% formic acid, 0.8M GuHCl) at 0 degrees C, and samples immediately frozen at –80 degrees C. until further processed as above. Data on the deuterated sample sets was acquired in a single automated 8-hour run, and subsequent data reduction performed on the DXMS data reduction software as previously described. Corrections for loss of deuterium-label by individual fragments during DXMS analysis (after "quench") were made through measurement of loss of deuterium from reference α-spectrin R1617 samples that had been equilibrium-exchange-deuterated under denaturing conditions, as previously described (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Hamuro, et al. J. Mol. Biol. 4:703- 714 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Woods-Jr. U.S. Patent No. 6,599,707 (2003), Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003). High Resolution, residue-specific determination of amide hydrogen exchange rates from DXMS data (HR-DXMS) was performed as described below.

[0288] Equipment configuration. The equipment configuration consisted of electrically-actuated high pressure switching valves (Rheodyne), connected to two position actuators from Tar Designs Inc., Pittsburgh, as described previously (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Hamuro, et al. J. Mol. Biol. 4:703- 714 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Woods-Jr. U.S. Patent No. 6,599,707 (2003), Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003) . A highly modified

Spectraphysics AS3000 autosampler, partially under external PC control, employed a robotic arm to lift the desired frozen sample from the sample well, then automatically and rapidly melted and injected the sample under precise temperature control (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Hamuro, et al. J. Mol. Biol. 4:703- 714 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003). The autosampler basin was further thermally insulated and all but 20 vial positions were filled with powdered dry ice sufficient to keep samples colder than -45° C for 18 hours. Four HPLC pumps (Shimadzu LC-10AD) were operated by a Shimadzu SCL-10A pump controller. One produced forward flow over the pepsin column, another backflushed the protease column after sample digestion (0.05 % aqueous TFA), and two delivered solvents to a downstream HPLC column for gradient elution (A: 0.05 % aqueous TFA; B; 80% acetonitrile, 20% water, 0.01% TFA; 1 x 50 mm C18 Vydac # 218MS5105, pH 2.3). Valves, tubing, columns and autosampler were contained within a refrigerator at 2.8° C, with pepsin and HPLC columns immersed in melting ice. The timing and sequence of operation of the DXMS apparatus fluidics were controlled by a personal computer running an in-house written LabView-based program, interfaced to solid-state relays (digital input/output boards, National Instruments), controlling pumps, valve actuators, and MS data acquisition (Hamuro, et al. J. Mol. Biol. 323:871- 881 2002, Hamuro, et al. J. Mol. Biol. 4:703- 714 2002, Woods-Jr., et al. Journal of Cellular Biochemistry 37:89-98 2001, Hamuro, et al. J. Mol. Biol. 327:1065- 1076 2003, Woods-Jr. U.S. Patent No. 6,599,707 (2003), Zawadzki, et al. Protein Sci 12:1980-90 2003, Englander, et al. Proc. Nat. Acad. Sci. 100:7057- 7062 2003) .

[0289] COREX Calculations of α-spectrin R1617 amide exchange rate protection factors. Fyrestar, operating the COREX algorithm, was installed on the Blue Horizon supercomputer at the San Diego Supercomputer Center, and run in a sparse Monte Carlo mode against the structural coordinates of the α-spectrin R1617 construct employing a COREX window size of 8 and a sampling of 1000 states per partition. This resulted in a sampling of 8000 states of the total 900 million possible. Hydrogen exchange rates and protection factors were calculated from the stability profile as previously described (Hilser, et al. Proteins 27:171-83

1997, Hilser, et al. J Mol Biol 262:756-72 1996, Hilser, et al. Proc Natl Acad Sci U S A 95:9903-8 1998, Hilser Methods Mol Biol 168:93-116 2001).

Algorithm and Software for High Resolution, residue-specific determination of amide hydrogen exchange rates from DXMS data: HR-DXMS

[0290] There does not currently exist an algorithm that will reliably find a globally minimum value for an arbitrary non-linear function. A common numerical difficulty in non-linear optimization is the discovery of local optima which exhibit many of the properties of globally minimal points, but are not in fact globally minimal. All currently known algorithms for non-linear optimization are susceptible to the problem of incorrectly terminating at a local minimum; however, in the case of smooth and continuous objective functions, one can often ameliorate this problem by initializing the numerical optimization with a solution that is likely to be near the global minimum.

[0291] The following algorithm centers on use of a two-phase numerical technique, linear programming (LP) followed by nonlinear least squares (NLS). The computational problem is to determine the mass gain ("shifts") for each smallest segment of the protein's sequence (here termed "atomic unit" or AU) that is resolved by differences between each DXMS-generated overlapping fragments' sequence, at each time point measured in the experiment. Linear regression of the AU shifts was applied to determine a rate for each AU. The rates from the linear regression analysis of the shifts are then fed into the nonlinear least-squares technique as initial rates.

[0292] Linear Programming Method: Linear programming (LP) is a technique that optimizes an objective function subject to certain predefined linear constraints. Given a protein sequence P where P_i denotes the *i-th* character of P, a fragment $f_{i,j}$ is simply a substring of P: $P_i, P_{i+1} \ldots P_j$. Fragments are generated by the protein digestion phase of the DXMS experiment, and are generally fixed for a given data analysis problem. A position k in the protein is covered by any fragment $f_{i,j}$ when $i \leq k \leq j$. An AU is the largest consecutive substring whose positions are covered by the same set of fragments, and can not overlap. The concatenation of all AU generated by the fragments in an experiment will cover each position of P only once if, and only if, P is the union of all the fragments. Therefore the set of AU is

determined entirely by the set of fragments generated by protein digestion. With the α-spectrin R1617 fragmentation map of 114 peptides there are 65 AU spanning 100% of the protein sequence with sizes ranging from 1 to 13 amides in length. By calculating the AU from the fragmentation pattern and knowing the mass shift of each AU at each time point one can calculate the mass shift, $d_{au}(t)$, for each AU as well as the corresponding error in our estimation. Note that if the fragmentation pattern produced 212 AU, that would represent single amide coverage over the entire 212 amino acid sequence of α-spectrin R1617.

[0293] Therefore, *A* is the set of AU determined by *F* (the set of fragments). For each fragment *f*, where *f* is a subset of *F*, there exists a set of of AU whose positions are covered by *f*. For each AU, *A(i)*, we define a variable $s_{i,t}$ that represents the mass shift of AU *i* at time *t*. For each fragment *f*, we define a variable $E_{f,t}$ which represents the experimental error in the mass shift measurement for fragment *f* at time *t*. The computational problem is to determine the mass gain ("shifts"), $s_{i,t}$, for each AU, *A(i)*, at each time point measured in the experiment. Figure 13A illustrates the definition of the AU for the first 15 amino acid segment of R1617. Atomic units (*A1, A2... A8)* are defined by the set of fragments *(f1, f2... f12)* and each fragment shift is the additive contribution of the calculated shifts for each AU, Figure 13B. After determining the mass shift for each AU at each time point linear regression of the AU's shifts is applied to determine a rate for each AU. The rates so calculated represent average rates of exchange of all amide hydrogens within the AU and provide good initial starting rates to seed into the non-linear least squares fit.

[0294] Non-Linear Least Squares Fit: The exchange process in a protein of *N* amino acids can be approximated as *N* independent chemical reactions that each obey first-order reaction kinetics. In particular, if amino acid *i* has rate constant $k_{ex,i}$ then the amount of deuterium $D_i(t)$, at time *t* at position *i* is simply

$$D_i(t)=1-e^{-k_{ex,i}\,t} \qquad 1.$$

The rate constant $k_{ex,i}$ is a function of pD, temperature, protein sequence, and protein conformation. For a fragment *f* composed of *n* amides the amount of deuterium incorporated is

$$D_{F(),t} = S^{n}_{i=m}(1-e^{-k_{ex,i}\,t}) \qquad 2.$$

where $D_{F(),t}$ is the total amount of deuterium on fragment f starting at amino acid residue m through amino acid residue n at time t, and $\mathbf{k_{ex,i}}$ is the exchange rate constant of amide i, where $m \leq i \leq n$. For the nonlinear least-squares technique the computational problem is to find rate constants that minimize the squared difference between the theoretical deuteration level of all measured fragments (equation 2) and the fragments' observed levels. The objective function is aimed to minimize the global error (GE) and includes a form of equation 2 for all fragments (p) at all time points (z) and attempts a global fit over all parameters according to a simplified equation 3 for our spectrin analysis of 114 fragments.

$$S^{p}_{f=1}[S^{z}_{t=1}D_{obs,F(),t} - D_{F(),t})^2\ ] = \text{GE} \qquad 3.$$

The parameters in this optimization ($k_{ex,i}$) are exactly the quantities of interest. Equation 3 can be used if the back exchange (loss of deuterium from protein/ peptides after the institution of "exchange-quench" conditions) of the peptides is corrected by using the standard peptide average exchange method. For a more rigorous correction of back exchange, one can correct for the loss of deuterium on each amide independently by modifying equation 3 to include published off exchange rates of model peptides (Molday, et al. Biochemistry 11:150-8. 1972, Bai, et al. Proteins 17:75-86. 1993).

[0295] In a completely unstructured polypeptide chain, all peptide amide hydrogens are freely accessible to water and exchange at their maximal possible rate, with a half-life of exchange of approximately one second at 0 °C and pH 7.0. Exact exchange rates for particular amide hydrogens in fully unstructured sequence can be reliably calculated from knowledge of the temperature, pH and primary amino acid sequence involved (Molday, et al. Biochemistry 11:150-8. 1972, Bai, et al. Proteins 17:75-86. 1993). The precise rate of exchange of a particular amide in random coil can vary more than thirty -fold from the average rates for all amides in a peptide under such conditions, with the precise rate depending upon the identity of the two amino acids flanking the particular amide bond, and whether or not the amide is at the c- or n- terminus of the peptide (Molday, et al. Biochemistry 11:150-8. 1972, Bai, et al. Proteins 17:75-86. 1993). The N- terminal amide in a peptide generally exchanges 20 times faster than the average rate for the other amides in

most peptides, a phenomenon that is important to take into account in data reduction calculations. Because DXMS analysis fragments and denatures peptides, one can model the off-exchange of amide deuterium from the fragments as a random coil and represent it as

$$D_{off,i}(T) = e^{-k_{int,i}(T_q)} \qquad 4.$$

where $D_{off,i}(T)$ is the fraction of deuterium left on a deuterated amide given an off-exchange time of T, and k_{int} is the intrinsic exchange rate of the amide under quench conditions (pH 2.3, 273K) calculated from known rates of model peptides from Bai and Englander (Molday, et al. Biochemistry 11:150-8. 1972, Bai, et al. Proteins 17:75-86. 1993). T represents the time upon quench to the time the fragment is analyzed in the mass spectrometer and is the sum of the fragment's retention time and the system lag time (SLT), the time between induction of exchange quench and sample loading onto the C18 column (2-5min). Although the retention times for each fragment are readily determined, the SLT can be better approximated to a value which results in the least amount of error in the overall fit, as described below.

[0296] Equation 2 for the total deuterium on a fragment can be readily modified to incorporate amide specific back exchange rate for every amide on that fragment by substitution of equation 4.

$$D_{corr,F(f),t} = S^{n}_{i=m} D_{off,i}(T)(1-e^{-k_{ex,i}t}) \qquad 5.$$

Now $D_{corr,F(f),t}$ is the corrected total amount of deuterium for fragment *f* at time *t* taking into account amide specific back-exchange rates which are dependent on the fragments retention time in the system under quench conditions. Substituting $D_{corr,F(f),t}$ into equation 3 allows one to refit the exchange rates with the corrections automatically taken into account

$$S^{p}_{f=1}[S^{z}_{t=1}D_{obs,F(f),t} - D_{corr,F(f),t})^2\ ] = GE \qquad 6.$$

[0297] Validation studies. The success of the method relies on the extent of overlapping fragmentation, the number of sampled on–exchange time points, and the number of post-quench off exchange time points sampled. Simulation studies were performed to determine the overall performance of this approach when using values for these parameters that were readily achievable in the present study.

[0298] α-Spectrin R1617 construct simulations. Studies were performed to determine how accurately the HR-DXMS method could deconvolute input DXMS data, with a fragmentation intensity and number of on-exchange time points similar to those employed with the α-spectrin R1617 construct in the present study. An arbitrary exchange rate map for a hypothetical "Hyp R1617 protein" was generated that was approximately based on the rates calculated for α-spectrin R1617 by COREX analysis (Figure 11C). Given these hypothetical rates, predicted deuteration levels for each of the same peptide fragment sequences (200 fragments) collected in the actual DXMS experiment with α-spectrin were generated, with and without incorporation of a normally-distributed random variable to simulate experimental and instrumental error. The HR-DXMS algorithm was used to simulate deuterated fragmentation data, and compared the resulting deconvoluted amide-specific determinations (expressed as the free energy of exchange; (Figure 14) to the free energy of exchange profile of the "Hyp R1617 protein" used to generate the data (Figure 14). There was a very good agreement between the two profiles when using simulated fragmentation data without error, which was only minimally degraded when a 20% error in peptide deuteration levels was added to the simulated data set prior to deconvolution. In the course of the present work, it has been observer that peptide deuteration levels are typically measured with a precision of 5-10% in DXMS, due in large part to the reproducibility resulting from the extensive automation employed. Regions where individual amides considerably diverged can be observed by the lack of overlap between the blue and lavender colored lines (Figure 14).

[0299] Horse cytochrome c. HR-DXMS was used to examine simulated DXMS deuterated fragment datasets based on published NMR-determined experimental hydrogen exchange rate data from horse cytochrome c (Milne, et al. Protein Sci 7:739-45. 1998). Residues where the rates of exchange had been too fast to be measurable in the NMR experiments were assigned arbitrary values. Since horse cytochrome c is 104 amino acids in length we used the same fragmentation pattern as that obtained for α-spectrin R1617 for the first 104 amino acid residues. Figure 15 shows that the experimentally-determined exchange amide-specific free energy profile of exchange of cytochrome c agree closely with the HR-DXMS- deconvoluted rate profile of the simulated data.

[0300] An important necessity to proper behavior of the fitting algorithm is the imposition of upper and lower bounds during the nonlinear least squares fit. Since the slowest exchanging peptides reached 50% deuteration level at 10^5 sec this corresponds to an average exchange rate on the order of 10^{-6}/sec. Lower boundaries were set 2 orders of magnitude lower so as to not exclude the possibility that a single amide may show slower rates within a given peptide, with the exception of regions of the protein sequence that had peptides that were maximally deuterated at the 10sec time point. With these peptides, the lower boundary of exchange was calculated at .92/sec, corresponding to 99.99% deuteration at 10 secs. The upper boundaries for the fit were set to the maximum exchange rates of each amide (Molday, et al. Biochemistry 11:150-8. 1972, Bai, et al. Proteins 17:75-86. 1993).

[0301] While the invention has been described and exemplified in sufficient detail for those skilled in this art to make and use it, various alternatives, modifications, and improvements should be apparent without departing from the spirit and scope of the invention. The present invention is well adapted to carry out the objects and obtain the ends and advantages mentioned, as well as those inherent therein.

[0302] The examples provided here are representative of preferred embodiments, are exemplary, and are not intended as limitations on the scope of the invention. Modifications therein and other uses will occur to those skilled in the art. These modifications are encompassed within the spirit of the invention.

[0303] The disclosure of all publications cited above are expressly incorporated herein by reference, each in its entirety, to the same extent as if each were incorporated by reference individually.